EXPLORING MOUSE AND HUMAN PLURIPOTENT STEM CELL FATE DETERMINATION
WITH BAYESIAN NETWORK MACHINE LEARNING

By
Karen Garvin Dowell
B.A. University of California, Berkeley, 1980
B.S. University of Southern Maine, 2008

A DISSERTATION
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
(in Biomedical Sciences)

The Graduate School
University of Maine
August 2013

Advisory Committee:
Matthew A. Hibbs, Assistant Professor, Trinity University, Co-Advisor
Gary A. Churchill, Professor, The Jackson Laboratory, Co-Advisor
Keith W. Hutchison, Professor, Committee Chair
Judith A. Blake, Associate Professor, The Jackson Laboratory
Carol J. Bult, Professor, The Jackson Laboratory
Zack Z. Wang, Director, Ross Flow Cytometry Core, Johns Hopkins Medicine
Kyunon Yun, Assistant Professor, The Jackson Laboratory
EXPLORING MOUSE AND HUMAN PLURIPOTENT STEM CELL FATE DETERMINATION

WITH BAYESIAN NETWORK MACHINE LEARNING

By

Karen Garvin Dowell

B.A. University of California, Berkeley, 1980

B.S. University of Southern Maine, 2008

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Biomedical Sciences)

The Graduate School

University of Maine

August 2013

Advisory Committee:

Matthew A. Hibbs, Assistant Professor, Trinity University, Co-Advisor
Gary A. Churchill, Professor, The Jackson Laboratory, Co-Advisor
Keith W. Hutchison, Professor, Committee Chair
Judith A. Blake, Associate Professor, The Jackson Laboratory
Carol J. Bult, Professor, The Jackson Laboratory
Zack Z. Wang, Director, Ross Flow Cytometry Core, Johns Hopkins Medicine
Kyu Son Yun, Assistant Professor, The Jackson Laboratory
On behalf of the Graduate Committee for Karen G. Dowell, we affirm that this manuscript is the final and accepted dissertation. Signatures of all committee members are on file with the Graduate School at the University of Maine, 42 Stodder Hall, Orono, Maine.

Dr. Matthew A. Hibbs, Assistant Professor    July 9, 2013

Dr. Gary A. Churchill, Professor           July 9, 2013
In presenting this dissertation in partial fulfillment of the requirements for an advanced degree at The University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for “fair use” copying of this dissertation for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this dissertation for financial gain shall not be allowed without my written permission.

Signature: _______________________________________________

Date: _______________________________________________
EXPLORING MOUSE AND HUMAN PLURIPOTENT STEM CELL FATE DETERMINATION WITH BAYESIAN NETWORK MACHINE LEARNING

By Karen G. Dowell

Dissertation Co-Advisors: Dr. Matthew A. Hibbs and Dr. Gary A. Churchill


Pluripotent stem cell research is an active, often controversial field focused on a special type of cell that exists in vivo only during the earliest stages of embryonic development. These unique cells have the ability to self-renew to maintain their stem cell population or differentiate into cells with the potential to become any cell type in the developing embryo. By investigating pluripotent stem cells in vitro, researchers gain insights into biological processes, pathways, and interactions that influence early development, and contribute to our ever-evolving understanding of cellular reprogramming, tissue regeneration, degenerative diseases, and cancers. Although intensively studied for decades, the molecular mechanisms that dictate pluripotent stem cell fate are diverse and remain only partially understood.

The objective of this study is to clarify the molecular foundations of mouse and human pluripotent stem cell self-renewal by identifying novel genes and interactions that influence self-renewal and contrasting self-renewal processes across species. For this work, I use a straightforward and flexible Bayesian network machine learning methodology for genomic data integration to generate consensus networks among protein-coding genes. To minimize
confounding factors, I focus on one type of pluripotent stem cell, embryonic stem cells (ESCs), in two species, mouse and human. This study provides intriguing insights into novel genes involved in mouse and human ESC self-renewal, highlights important differences in developmental signaling and metabolic pathways that support ESC self-renewal in these species, and serves as the foundation for future pluripotent stem cell and cancer stem-like cell studies. To make the results of this study accessible to the research community, I provide a comprehensive set of computational resources (available online at www.StemSight.org) that includes a powerful, dynamic network visualization tool designed to enable biologist end-users to explore my dense, complex biological networks. These resources may be used by stem cell researchers to discover novel regulators of embryonic stem cell self-renewal, test new hypotheses, and prioritize genes of interest for further study.
DEDICATION

For my father, Frank Dowell, and in memory of my mother, Marian Dowell, both of whom inspired me to never stop learning.
ACKNOWLEDGMENTS

I was fortunate to have several passionate teachers at the University of Southern Maine who nurtured my interest in developmental biology, comparative genomics, and the biomedical sciences. David Champlin encouraged me to sign up for his graduate-level seminar on functional genomics and later served as my academic sponsor for a work study project in Doug Spicer’s laboratory at Maine Medical Center Research Institute. Tom Knight, who blended teaching with his own special brand of storytelling, introduced me to molecular biology in the genomics age. I would not have applied to the University of Maine Graduate School of Biomedical Sciences and Engineering (GSBSE) were it not for Dave, Doug, and Tom.

I am grateful to all the members of my advisory committee for their invaluable feedback and advice during my predoctoral career: Matt Hibbs, Gary Churchill, Keith Hutchison, Judith Blake, Carol Bult, Zack Wang, and Kyuson Yun. If it takes a village to raise a child, it took three research groups to shepherd me through this program, and I benefited from the advice, support, and friendship of many current and former members of the Blake, Hibbs, and Churchill labs, especially Harold Drabkin, David Hill, Monica McAndrews, Al Simons, Joe Bane, Tongjun Gu, Cheryl Ackert-Bicknell, Kwangbum “KB” Choi, Imogen Hurley, Dan Gatti, Sue McClatchy, and Narayanan Raghupathy. The Jackson Laboratory (JAX) was an extraordinary research setting to pursue a PhD, and I thank those responsible for the JAX predoctoral education initiative while I was in residence: Mary Ann Handel, Suzanne Serreze, and Michael McKernan. In addition, I thank Carol Kim, Keith Hutchison and Laura Hall from the GSBSE, as well as Carol Bult and Professor Emeritus Barbara Knowles of JAX, without whom this multi-institutional, functional genomics program might not exist. Most of all, I thank my primary mentor, Matt Hibbs. Matt is
of a new generation of computational biologists who are changing the way scientific researchers view, use, and think about biological information. I was fortunate to have been his first graduate student and a member of his computational biology and functional genomics research team at JAX. One of the best teachers I have ever had, Matt introduced me to machine learning with big data and set me on this computational biology path. His advice and direction throughout my graduate student career have been invaluable.

I thank my family, friends, and fellow students for their support, encouragement, and patience while I worked on this project. I am especially grateful to my father, Frank Dowell, who has always been my role model of erudition; Mark Billings, who cheered me on while gently reminding me that balance in life is important; and my beloved dogs, Harry, Mitchell and Seager, who spent much of their lives at my feet, waiting for me to finish this project.

Chapter 2 is joint work with Al Simons, Zack Wang, Kyuson Yun, and Matt Hibbs. It has been published in PLoS One [1]. Al made many technical contributions to this project and wrote several C++ and Python scripts that enabled me to more efficiently preprocess data and computationally validate networks (he was also an excellent programming teacher). Zack and Kyuson served as stem cell experts, and helped verify that my predictive networks were biologically relevant and helped position results for the stem cell biologist audience. Chapter 3 is joint work with Al Simons, Braden Kell, and Matt Hibbs; a version of this work will be submitted to the journal Bioinformatics. Chapter 4 is joint work with Al Simons, Braden Kell, Zack Wang, Kyuson Yun, and Matt Hibbs, and will be submitted to the journal Stem Cells.

This work was funded by The Jackson Laboratory, a Pharmaceutical Researchers and Manufacturers of America (PhRMA) Foundation Fellowship in Informatics, University of Maine Graduate School of Biomedical Sciences and Engineering, and the National Science Foundation Integrated Graduate Education and Research Traineeship (IGERT) program grant 0221625.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ...................................................................................................................... v

LIST OF TABLES ........................................................................................................................ xii

LIST OF FIGURES ....................................................................................................................... xiv

LIST OF EQUATIONS ................................................................................................................... xvi

CHAPTER 1: INTRODUCTION .......................................................................................................... 1

Understanding the Potential of Pluripotent Stem Cells .......................................................... 2

Deciphering the Molecular Basis of Stem Cell Self-Renewal ................................................ 3

Self-Renewal in Pluripotent Stem Cells Caters to Growth-Intensive Embryonic Development ................................................................. 5

Self-Renewal in Lineage-Restricted Stem Cells Conforms to Changing Needs of Tissues ................................................................. 6

Reprogrammed Adult Cells Behave like ESCs ........................................................................ 8

Stem-Like Cancer Cells May Hijack Stem Cell Self-Renewal Machinery .......................... 9

Systems Biology Knowledge Discovery and Data Mining Techniques .............................. 10

A Systems Approach to Investigating ESC Self-Renewal ................................................... 13

Inferring Biological Networks from Diverse Functional Genomic Data ............................ 15

Applying Bayesian Network Machine Learning Approaches to Mammalian Systems .................. 17

Bayesian Network Machine Learning for Protein Function Prediction in Pluripotent Stem Cells .................................................................................................................. 19

Contributions and Project Overview ....................................................................................... 19
CHAPTER 2: CELL-TYPE-SPECIFIC NETWORK YIELDS INSIGHTS INTO MOUSE EMBRYONIC STEM CELL SELF-RENEWAL AND CELL FATE ................................................................. 21

Abstract ................................................................................................................................. 21

Introduction .............................................................................................................................. 22

Computational Methods for Predicting Protein Function Using High-Throughput Data ............ 23

Results ................................................................................................................................... 27

mESC-Specific Network Predicts Novel Self-Renewal Proteins ............................................ 27

Appropriate Evidential Data is Critical for Useful Network Predictions ................................. 38

Computational Performance Metrics Do Not Necessarily Measure Biological Relevance ......... 42

Network Visualization Reveals Novel Functional Relationships .......................................... 44

Discussion ............................................................................................................................... 45

Statistical Analysis of Input Data Enhances Relevance of Biological Networks .......... 45

Tailoring Prior Knowledge to a Single Cell-Type Clarifies Results .......................................... 50

Cellular Context Matters in Predictive Biological Networks ................................................. 52

Conclusions ............................................................................................................................ 53

Materials and Methods .......................................................................................................... 55

Collection and Preparation of Training and Evidential Stem Cell Knowledge ...................... 55

Preparation of Evidential High-Throughput Mouse Embryonic Stem Cell Dataset Compendium ................................................................................................................. 56

Prepare Curated, Tissue-Specific Training Gold Standard ....................................................... 59

Construct Bayesian Network and Infer Posterior Functional Relationship Scores............ 59

Minimization of Network Overfitting ...................................................................................... 60
CHAPTER 3: STEMSIGHT SCOUT: DYNAMIC DATA VISUALIZATION FOR PLURIPOTENT STEM CELL DATA ANALYSIS AND GENE FUNCTION DISCOVERY

Abstract........................................................................................................................................... 68
Introduction....................................................................................................................................... 68
Implementation................................................................................................................................. 69
Available Features........................................................................................................................... 70
Tutorial: Using Scout to Explore Predictive Networks............................................................ 75
Creating and Refining Gene-Centered Networks................................................................. 76
Creating Multi-Centered Gene Networks.............................................................................. 78
Collecting Network Details for Further Analysis and Reports ........................................ 79
Future Plans.................................................................................................................................... 82

CHAPTER 4: NOVEL INSIGHTS INTO EMBRYONIC STEM CELL SELF-RENEWAL REVEALED THROUGH COMPARATIVE HUMAN AND MOUSE SYSTEMS BIOLOGY NETWORKS

Abstract............................................................................................................................................ 83
Introduction....................................................................................................................................... 84
APPENDICES

Appendix A. Glossary of Terms and Abbreviations................................................................. 130
Appendix B. Supplemental Figures ....................................................................................... 136
Appendix C. Supplemental Notes ......................................................................................... 146
Appendix D. Supplemental Materials and Methods .............................................................. 162
Appendix E. Supplemental Tables and Image Files ............................................................... POCKET

BIOGRAPHY OF THE AUTHOR ............................................................................................ 174
LIST OF TABLES

Table 2.1. Summary of Integrated mESC Genomic Data .......................................................... 31
Table 2.2. Candidate Genes for Experimental Validation ...................................................... 35
Table 2.3. Comparison of mESC-Specific and Test Network Connectivity ............................. 44
Table 4.1. Summary of Integrated Data in Human and Mouse ESC Data Compendiums ...... 89
Table 4.2. Top Novel hESC Genes Most Strongly Correlated with Self-Renewal ................ 95
Table 4.3. Functional Similarities and Differences Between Networks ................................. 97
Table P.1. mESC Self-Renewal Gold Standard Positive Edges and Supporting Literature References ........................................................................................................ POCKET
Table P.2. Journal Articles Referenced to Curate mESC Gold Standard ............................... POCKET
Table P.3. mESC Data Compendium .................................................................................... POCKET
Table P.4. mESC Network Gene Ontology Annotation Enrichment .................................... POCKET
Table P.5. Minimalist Data Compendium ............................................................................. POCKET
Table P.6. Negative Control Data Compendium .................................................................... POCKET
Table P.7. Superset Data Compendium ................................................................................. POCKET
Table P.8. Data Compendium Tests ....................................................................................... POCKET
Table P.9. Network Topology Statistics for mESC and Mmu Networks ............................... POCKET
Table P.10. Gold Standard Posterior Edge Weight and Rank Order ...................................... POCKET
Table P.11. Datasets Supporting Top Ranked mESC Network Edges ................................... POCKET
Table P.12. Protein Coding Gene List .................................................................................... POCKET
Table P.13. mESC Gold Standard Answer File ..................................................................... POCKET
Table P.14. Protein Coding Gene Ortholog List .................................................................... POCKET
Table P.15. ESC Self-Renewal Gold Standard Positive Edges and Supporting Literature References ........................................................................................................ POCKET
Table P.16.   Journal Articles Referenced to Curate ESC Gold Standard ............................... POCKET

Table P.17.   ESC Gold Standard Answer File (Positive and Negative Training
Examples) .................................................................................................................. POCKET

Table P.18.   hESC Data Compendium ........................................................................ POCKET

Table P.19.   mESC Data Compendium ........................................................................ POCKET

Table P.20.   Network Gene Ontology Annotation Enrichment ................................... POCKET

Table P.21.   Comparative Network Topology Statistics ............................................. POCKET

Table P.22.   Gold Standard Posterior Edge Weight and Rank Order.......................... POCKET

Table P.23.   Differential Network Analysis ...................................................................... POCKET
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Understanding Stem Cell Self-Renewal.</td>
<td>4</td>
</tr>
<tr>
<td>1.2</td>
<td>Stem Cell Developmental Ontogenies</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>Basic Steps of Biological Systems Analysis Workflows</td>
<td>11</td>
</tr>
<tr>
<td>1.4</td>
<td>Naïve Bayesian Networks for Genomic Data Integration</td>
<td>16</td>
</tr>
<tr>
<td>2.1</td>
<td>Naïve Bayesian Networks for Genomic Data Integration</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>Cell-Type-Specific Data Integration and Machine-Learning Methodology.</td>
<td>29</td>
</tr>
<tr>
<td>2.3</td>
<td>Network Performance Evaluations</td>
<td>32</td>
</tr>
<tr>
<td>2.4</td>
<td>Data Visualization for Mining mESC Self-Renewal Gene Predictions</td>
<td>37</td>
</tr>
<tr>
<td>2.5</td>
<td>Importance of Feature Selection in Bayesian Network Machine Learning</td>
<td>40</td>
</tr>
<tr>
<td>2.6</td>
<td>Advantages of a Statistically Principled Approach</td>
<td>47</td>
</tr>
<tr>
<td>2.7</td>
<td>Comparing Subnetworks of WNT Signaling Pathway Participants.</td>
<td>48</td>
</tr>
<tr>
<td>3.1</td>
<td>StemSight Scout User Interface</td>
<td>72</td>
</tr>
<tr>
<td>3.2</td>
<td>Using StemSight Scout to Explore Network Predictions</td>
<td>80</td>
</tr>
<tr>
<td>4.1</td>
<td>Bayes Net Machine Learning Methodology for Cell-Type-Specific Comparative Networks</td>
<td>93</td>
</tr>
<tr>
<td>4.2</td>
<td>Conservation of Self-Renewal Functional Associations Across Species</td>
<td>98</td>
</tr>
<tr>
<td>4.3</td>
<td>Predicted FGF Signaling Pathway Relationships Across Species</td>
<td>101</td>
</tr>
<tr>
<td>4.4</td>
<td>Predicted JAK/STAT Signaling Pathway Relationships Across Species</td>
<td>103</td>
</tr>
<tr>
<td>4.5</td>
<td>Differences in mESC and hESC Threonine Metabolism</td>
<td>108</td>
</tr>
<tr>
<td>B.1</td>
<td>mESC Network Performance Evaluation through Cross Validation</td>
<td>135</td>
</tr>
<tr>
<td>B.2</td>
<td>Regularization and Self-Renewal Connectivity</td>
<td>137</td>
</tr>
<tr>
<td>B.3</td>
<td>Dataset Classes Supporting Top mESC Network Edges</td>
<td>139</td>
</tr>
<tr>
<td>B.4</td>
<td>Data Compendium Tests</td>
<td>140</td>
</tr>
</tbody>
</table>
Figure B.5. Datasets Supporting Tdh Connectivity to Gold Standard Genes ...................... 142
Figure B.6. Network Performance Evaluations ................................................................... 143

Note: High-resolution image files of figures in Chapters 2–4 and Appendix B are provided in Appendix E, which is located on the Pocket CD included with this dissertation.
<table>
<thead>
<tr>
<th>Equation</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Pearson Correlation</td>
<td>57</td>
</tr>
<tr>
<td>2.2</td>
<td>Fisher’s Z Transformation</td>
<td>57</td>
</tr>
<tr>
<td>2.3</td>
<td>Dot Product Calculation</td>
<td>58</td>
</tr>
<tr>
<td>2.4</td>
<td>Euclidean Distance Measure</td>
<td>58</td>
</tr>
<tr>
<td>2.5</td>
<td>Inference of Posterior Functional Relationship</td>
<td>60</td>
</tr>
<tr>
<td>2.6</td>
<td>Heuristic Sum of Shared Information Relative to Dataset Entropy</td>
<td>63</td>
</tr>
<tr>
<td>2.7</td>
<td>Inference of Posterior Functional Relationship with Regularization</td>
<td>63</td>
</tr>
<tr>
<td>2.8</td>
<td>True Positive Rate</td>
<td>64</td>
</tr>
<tr>
<td>2.9</td>
<td>False Positive Rate</td>
<td>64</td>
</tr>
<tr>
<td>2.10</td>
<td>Precision</td>
<td>64</td>
</tr>
<tr>
<td>2.11</td>
<td>Recall</td>
<td>64</td>
</tr>
<tr>
<td>2.12</td>
<td>Degree</td>
<td>65</td>
</tr>
<tr>
<td>2.13</td>
<td>Sum of Degrees</td>
<td>65</td>
</tr>
<tr>
<td>2.14</td>
<td>Mean Degree</td>
<td>65</td>
</tr>
<tr>
<td>2.15</td>
<td>Scaled Degree</td>
<td>65</td>
</tr>
<tr>
<td>2.16</td>
<td>Functional Correlation Score</td>
<td>66</td>
</tr>
<tr>
<td>4.1</td>
<td>Inference of Posterior Functional Relationship</td>
<td>90</td>
</tr>
<tr>
<td>4.2</td>
<td>Heuristic Sum of Shared Information Relative to Dataset Entropy</td>
<td>90</td>
</tr>
<tr>
<td>4.3</td>
<td>Inference of Posterior Functional Relationship with Regularization</td>
<td>90</td>
</tr>
<tr>
<td>D.1</td>
<td>Pearson Correlation</td>
<td>163</td>
</tr>
<tr>
<td>D.2</td>
<td>Fisher’s Z Transformation</td>
<td>163</td>
</tr>
<tr>
<td>D.3</td>
<td>Euclidean Distance Measure</td>
<td>164</td>
</tr>
</tbody>
</table>
Equation D.4. Dot Product Calculation................................................................. 164
Equation D.5. Inference of Posterior Functional Relationship............................... 167
Equation D.6 Heuristic Sum of Shared Information Relative to Dataset Entropy........ 168
Equation D.7. Inference of Posterior Functional Relationship with Regularization........ 168
Equation D.8. True Positive Rate........................................................................... 170
Equation D.9. False Positive Rate........................................................................... 170
Equation D.10. Precision .................................................................................... 170
Equation D.11. Recall ......................................................................................... 170
Equation D.12. Degree ....................................................................................... 171
Equation D.13. Sum of Degrees........................................................................... 171
Equation D.14. Mean Degree............................................................................... 171
Equation D.15. Scaled Degree............................................................................. 171
Equation D.16. Functional Correlation Score....................................................... 171
Equation D.17. Spearman Correlation Coefficient Between Ranked Variables......... 173
Equation D.18. Fisher’s Z Transformation............................................................. 173
Equation D.19. Z-Score Calculation....................................................................... 173
CHAPTER 1

INTRODUCTION

Stem cells have long garnered the interest and excitement of the biomedical community. These special cells have the capacity to divide and renew their cell populations indefinitely and produce differentiated cell types that perform more specialized functions. Since Till and McCulloch first proposed the concept of a cellular hierarchy in which multiple cell types are derived from a single population of “self-renewing” stem cells in 1963, we have amassed a storehouse of knowledge about these cells and their unique properties [2, 3]. Despite continuing and intensive research efforts in the field, the underlying molecular mechanisms that enable stem cells to both self-renew and produce differentiated progeny remain elusive. This is largely because there is not one cell fate pathway, but rather several variations on a pathway theme, contingent on the type of stem cell, its niche, epigenetic state, and myriad internal and external signaling cues [4, 5].

The objective of this study is to clarify the molecular foundations of stem cell self-renewal by identifying novel genes and interactions that influence self-renewal and revealing differences in self-renewal processes across species. For this work, I use a straightforward and flexible Bayesian network machine learning methodology for genomic data integration to generate consensus networks that predict functional associations among protein-coding genes. To minimize confounding factors, I focus on self-renewal in one type of stem cell, embryonic stem cells (ESCs), in two species: mouse and human. I chose this stem cell type because the origin or lineage of ESCs in vivo is well understood, whereas the origin of lineage-specific adult stem cells is less clear and, in some cases, controversial [6]. Mouse and human ESCs (mESCs and
hESCs, respectively) have been exhaustively studied; they are well-documented in the literature, and extensive high-throughput data for these cells are available in public repositories.

This study provides intriguing insights into novel genes involved in mouse and human ESC self-renewal, and includes a comprehensive set of computational resources designed to help stem cell researchers more effectively explore known and novel functional associations among protein-coding genes in the context of ESC self-renewal in mouse and humans, and serves as a roadmap for future self-renewal studies in other stem cell types. Supporting materials and tools mentioned in this dissertation are available online at www.stemsight.org, and may be used by stem cell researchers to identify self-renewal genes, test new hypotheses, and prioritize genes of interest for further study. Collectively, these results enrich our understanding of comparative developmental biology, tissue regeneration, degenerative diseases, and cancers.

Understanding the Potential of Pluripotent Stem Cells

Pluripotent stem cell research is focused on the properties of cells that form during the earliest stages of embryonic development. During this period of rapid cell division and growth, the fertilized egg (or zygote) is cleaved into two, then four, then eight smaller cells, each with the same amount of zygotic cytoplasmic material. These cells, called blastomeres, are totipotent, meaning they have the potential to divide into any type of cell in the developing organism: the embryo and the fetal placenta or yolk sac that surrounds it. They cluster together in a sphere-like colony called a morula. As mitotic divisions continue, the blastomeres begin to segregate and position themselves to take on different functions. At approximately 100 cells (embryonic day 3.5-4.5 in mice, day 5-7 in humans), a hollow ball-like shape called a blastocyst
forms. It consists of a ring of outer cells (trophoblasts) that will become the fetal placenta or yolk sac, and an interior cluster of 10–20 cells, called the inner cell mass (ICM) that will develop into the embryo. The cells that comprise this ICM are pluripotent, meaning they have the potential to differentiate into any cell type in the developing embryo [7].

Pluripotent stem cells have been the subject of intense research since 1981, when Martin Evans and Matthew Kauffman at the University of Cambridge, and Gail Martin of the University of California at San Francisco, independently determined how to extract cells from the ICM of a mouse blastocyst and successfully grow them in culture [8, 9]. Dr. Martin coined the term “embryonic stem cells” (ESCs) for these embryo-derived pluripotent cells [9]. In 1998, the field of pluripotent research was further energized (and politicized) when James Thomson at the University of Wisconsin-Madison successfully isolated and cultured embryonic stem cells from human embryos [10]. Mouse and human ESCs have since become a powerful model system for developmental biologists and biomedical researchers interested in regenerative medicine, degenerative diseases, and cancers. By studying ESCs in vitro, researchers have learned how to direct stem cell fate in a dish [11], reprogram adult cells to take on characteristics of pluripotent stem cells [12], and test how cancer cells with stem-cell-like properties may self-renew and differentiate to produce tumors [13].

Deciphering the Molecular Basis of Stem Cell Self-Renewal

Self-renewal is a process by which stem cells divide to generate at least one daughter cell with the same developmental potential as the mother cell. This ensures that stem cells can proliferate during development, maintain a constant stem population within adult tissues, and repopulate the stem cell pool after injury. All stem cells are characterized by their ability to self-
renew and give rise to daughter cells with more restricted developmental potential [4-6, 14, 15]. For example, pluripotent ESCs beget lineage-specific multipotent cells, which, in turn, produce one or more types of terminally differentiated mature cells. Stem cells can divide symmetrically to generate two identical daughter cells or asymmetrically to produce one stem cell and one restricted progenitor cell (Figure 1.1) [4, 14].

The choreography of signaling required for these different flavors of self-renewal and differentiation are influenced by the developmental potential and lineage of the stem cell, the type of cell division, and the temporal and physical context (e.g. stage of development, tissue compartment, microenvironmental signals, etc.). Much of what we know about self-renewal mechanisms results from the study of specific types of stem cells in these different contexts, as well as in the synthetic stem cell niche that is tissue culture.

Figure 1.1. Understanding Stem Cell Self-Renewal. Stem cells have the capacity to self-renew symmetrically or asymmetrically.
Self-Renewal in Pluripotent Stem Cells Caters to Growth-Intensive Embryonic Development

Pluripotent ESCs possess almost unlimited capacity for self-renewal and can differentiate into all three germ layers in the developing embryo (endoderm, mesoderm, and ectoderm) and thus give rise to any cell types in the body [6, 7, 16]. In both human and mouse pluripotent cells, the transcriptional regulatory network trinity of Pou5f1 (also called Oct4), Sox2, and Nanog promotes self-renewal by inhibiting differentiation. These master regulators cooperatively activate or silence hundreds of genes, resulting in a gene expression and signaling profile that maintains the ESC state [5, 17]. The promoter regions of many of the genes silenced by these master regulators are also occupied by the Polycomb group (PcG) proteins, which are epigenetic regulators that catalyze histone methylation and facilitate the structural condensation of chromatin, thereby further hindering differentiation [17]. Developmental signaling pathways influenced by this cooperative regulation of gene expression include Notch, BMP, MAPK, JAK/STAT, Hedgehog, Wnt, FGF, and TGF-β [4, 5, 17-19]. (Appendix A lists full names for signaling pathway acronyms referenced in this document.) Recent studies have also shown that microRNAs (miRNAs), such as the Lethal-7 (Let-7) and miR-200 miRNA families, are involved in regulating ESCs and the opposing states of self-renewal and differentiation [20-23]. In addition, ESC self-renewal processes are facilitated by unusual cell cycle kinetics, largely driven by modifications to the P16^ink4a/CDK/Rb cell cycle pathway. These customized controls result in a clipped G₁ phase, rapid entry into the S phase, and fewer cell cycle checkpoints overall [5, 15]. Traditional models of ESC self-renewal tend to oversimplify highly complex mechanisms or focus only on a small subset of this multifaceted regulatory circuitry.

Mouse epiblast stem cells (mEpiSCs) are pluripotent stem cells isolated from the mouse epiblast, a cellular structure that forms post implantation, just prior to gastrulation [24, 25]. Developmentally and functionally distinct from mESCs, mEpiSCs are believed to be more closely
related to human ESCs in that they are “primed” to differentiate, whereas mESCs are in a more primitive, naïve state [26]. Unlike mESCs, female mEpiSCs have undergone inactivation of one X chromosome, and respond differently to developmental signaling pathways. In cell culture, mEpiSCs, similar to hESCs, require ACTIVIN-A-induced FGF signaling as opposed to LIF-induced JAK/STAT3 signaling to support self-renewal and thrive in culture. By contrasting predictive networks for these pluripotent stem cell types (Chapter 4), this study offers new insights into the epigenetic and signaling shifts that dampen self-renewal and trigger differentiation into ectoderm, mesoderm, and endoderm germ layers [25, 27-29].

Self-Renewal in Lineage-Restricted Stem Cells Conforms to Changing Needs of Tissues

Tissue-specific stem cells, such as hematopoietic stem cells (HSCs), mesenchymal stem cells (MSCs), and neural stem cells (NSCs), all have extensive capacity for self-renewal, but limited cell fate potential. These cells are multipotent, meaning they can give rise to multiple types of differentiated daughter cells within a distinct cell lineage (Figure 1.2) [5-7]. Multipotent stem cells and unipotent progenitor cells are shaped by their developmental path and subsequent niche. However, significant gaps remain in our knowledge of the cell fate triggers that induce differentiation within adult stem cell pathways and how these triggers may affect closely intertwined self-renewal mechanisms. It is not yet known at precisely which point in the development cycle tissue stem cells arise, nor how rapidly changing spatial dynamics and signaling stoichiometry within the inner cell mass of the blastula may influence multipotent cell fate [6, 7].
Figure 1.2. Stem Cell Developmental Ontogenies. There are varying views on how a single cell, the fertilized egg, can differentiate into all the cell types in the developing embryo (e.g. epidermal cells of skin, neurons, blood cells, bone tissue, cells of the digestive track, germ cells, etc.) and produce lineage specific stem cells, such as neural stem cells (NSCs), hematopoietic stem cells (HSCs), mesenchymal stem cells (MSCs), that are responsible for replenishing pools of adult cells throughout the life of the organism [6, 7]. These two alternative stem cell developmental ontogenies illustrate different possible scenarios leading to lineage-specific stem cells. (Adapted from Lanza et al.)
Mechanistically, multipotent (lineage-specific) stem cell self-renewal is distinct from the cell proliferation processes which maintain populations of developmentally restricted progenitor cells and committed adult cells. Unlike ESCs, these somatic stem cells can play very different developmental and homeostatic roles throughout the life of the organism and therefore require much finer cell cycle controls that allow for periods of cycling and quiescence [5-7, 30]. As in pluripotent stem cells, adult stem cell self-renewal programs are designed to halt differentiation and maintain stem cell populations. However, more specialized tissue-specific requirements necessitate tailored self-renewal programs that are strongly influenced by the stem cell niche. For example, NSCs require Bmi1 and Hmgα2 for self-renewal, whereas committed neuronal or glial progenitor cells do not [5, 31]. Similarly, Mll1 is required for HSC maintenance, but not for self-renewal of committed lymphoid or myloid progenitor cells [5, 30, 32]. Sox17 is required for maintenance of fetal and neonatal HSCs, yet it is not expressed in dividing HSCs, nor can it be induced in adult HSCs by proliferation or mobilization assays [5, 33].

Reprogrammed Adult Cells Behave like ESCs

Artificially induced pluripotent stem (iPS) cells, which are adult cells that have been “reprogrammed” to a pluripotent ESC-like state, are an promising experimental and therapeutic alternative to politically charged hESCs, but lingering issues with regard to reprogramming efficiency, efficacy, and safety limit practical biomedical applications [17, 34]. Transfecting adult cells with Yamanaka factors (i.e. Pou5f1, Sox2, Klf4, and Myc) has been shown to reprogram terminally differentiated cells, such as fibroblasts, into iPSCs that are molecularly and functionally similar to pluripotent embryonic cells [16, 35-37]. Yamanaka factors are also highly expressed in ESCs, suggesting that one or more of these factors may be key to maintaining the ESC pluripotent state. One study emphasized that Yamanaka factors collectively regulate as
many as 16 development signaling pathways, 12 of which are believed to be crucial to the pluripotent state of ESCs and iPS cells [36]. ESCs and iPS cells are also believed to share similar, poorly regulated cell cycle mechanisms [5, 17]. Several strategies for reprogramming somatic cells to a pluripotent state exist, each involving different cocktails of factors, delivery systems, and somatic cell types. In addition, there is currently little standardization among labs experimenting with iPS cells. It is not yet known how these variations and lack of standards will affect the integrity and therapeutic potential of these ES-like cells [17].

**Stem-Like Cancer Cells May Hijack Stem Cell Self-Renewal Machinery**

Although controversial, numerous recent studies have suggested the existence of cancer stem cells (CSCs), which are cancer cells that appear to possess stem-like traits. CSCs are thought to initiate tumors through unchecked proliferation. They maintain and extend their CSC population through self-renewal, and differentiate into mature cancer cells with restricted potential [15, 38-41]. As evidence of CSCs has grown, there has been increasing interest in exploring parallels between CSCs and stem cell populations to determine whether CSCs are mutated adult stem cells or terminally differentiated cells that have hijacked stem cell signaling pathways. Studies have shown that canonical stem cell factors *Pou5f1, Sox2*, and *Nanog* are overexpressed (in combination with the oncogene *Myc*) in poorly differentiated tumors [39, 42]. Interestingly, translocations in transcription factor genes associated with HSC self-renewal processes have been found to be associated with multiple leukemias, suggesting a strong link between deregulated self-renewal processes and malignant cancers [30, 43].

The type of self-renewal-focused systems analyses conducted in this work may reveal mechanistic parallels that will help elucidate the origin of tumor-initiating cancer stem cells and
provide evidence to support novel therapeutic approaches for “stemming” uncontrolled cancer cell growth.

**Systems Biology Knowledge Discovery and Data Mining Techniques**

Most existing analyses of stem cell self-renewal processes focus on the roles of individual genes or a small subset of genes, rather than examining genome-wide functional relationships and interactions between congregations of protein-coding genes and gene products. This reductionist approach tends to neglect how complex biological systems are influenced by functionally diverse, often multifunctional elements (such as genes, proteins, metabolites, miRNA, etc.), as well as the system response to activate, inhibit, produce or degrade these elements [44]. To build a true biological regulatory network requires a list of molecular elements and a set of regulatory interactions between these elements. In biological systems analyses (Figure 1.3), the list of elements typically comes from data derived from high-throughput experiments, such as microarray datasets, arrays based on chromatin immunoprecipitation (ChIP) pull-down methods followed by microarray (ChIP-chip) or sequencing (ChIP-seq), and whole genome RNA interference (RNAi) screens. Physical interaction data can be compiled from experiments applying techniques such as affinity purification followed by mass spectrometry (AP-MS). Functional connections between elements within a network can be statistically inferred by comparing experimentally derived expression patterns with curated interaction information extracted from the literature [45]. Combined, these types of techniques provide a more complete picture of the factors that influence one or more functionally related biological processes [44-46].
**Figure 1.3.** Basic Steps of Biological Systems Analysis Workflows. Most systems biology methods use the same fundamental approach consisting of a research and development phase followed by iterative modeling and exploration of the results [44-46].

Several recent studies have shown how computational knowledge discovery and data mining techniques can be used to extract hidden patterns from vast quantities of experimental data. When coupled with more traditional experimental approaches, systems analysis techniques can broaden our understanding of stem cell self-renewal processes.

For example, Wang *et al.* developed a protein interaction network for investigating pluripotency in mouse embryonic stem cells. Using a blend of affinity purification and proteomics, they identified proteins functionally associated with Nanog, and performed iterative screens to validate and test the functional relevance of these proteins. In this way, the authors were able to create a Nanog “interactome,” which revealed transcriptional patterns that suggest Nanog is involved in several protein complexes. The results of their analysis also
indicated that stoichiometry of individual elements within these complexes is important for maintaining pluripotency in mESCs [18]. Kim et al., elaborated on these findings by using biotinylation-mediated ChIP combined with microarray technology (BioChIP-Chip) to perform global transcriptional target mapping of an expanded set of factors associated with pluripotency in mESCs. They identified nine key components (Pouf51, Sox2, Nanog, Klf4, NrOb1, Nacc1, Zfp281, Zfp42, and Myc) of a transcriptional regulatory network involved in ESC pluripotency control and differentiation [47]. These types of studies illustrate how techniques, such as AP-MS and BioChIP-Chip, can be combined with microarray data to map a physical interaction network for mESCs. However, they show only one small piece of the overall pluripotency puzzle.

In another systems-based study, Wong et al. used a molecular module mapping technique to compare conditional expression activity in different types of stem cells, including ESCs, HSCs and NSCs, and cells from liver, breast, lung, and gastric tumors. They developed expression profiles based on 102 microarray data sets (representing more than 8 stem cell types). To categorize this data, they compiled ~3000 gene sets: lists of genes associated with specific biological processes as defined by the Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, and selected literature references [48, 49]. By contrasting stem cell expression profiles against these functional gene sets, they identified clusters of expression patterns, called modules, which they used to identify genes that act cooperatively to perform a biological function or contribute to a certain phenotype. They generated module heat maps to relate transcription programs among different types of stem cells, then compared the results to previously published module maps of cancer cells [50]. The authors’ analysis of these module maps revealed some shared, higher-order relationships among stem cell and stem-cell-like transcriptional programs, suggesting that many of the core ESC regulatory genes, developmental signaling pathways, and modified cell cycle regulation may be inappropriately turned on in
cancer cells [42]. This work illustrates how systems-level methods can be applied to investigate and compare transcription programs underlying self-renewal processes in stem cells. It could have been even more powerful had they started with a larger compendium of mouse stem cell data and extended their analysis beyond the module map. Other Segal Laboratory investigations using this approach to compare conditional activity of gene expression modules in different tumor types provide far more detailed evaluation of how gene expression signatures relate to specific known pathways extracted from the KEGG pathway database or from the literature [42, 50].

Studies such as these are extremely useful and provide clues to help deconstruct the mechanisms that drive fundamental stem cell biological processes, such as self-renewal and differentiation. But most of the predictive work done to date focused on pluripotent stem cells has been limited in scope and falls short of realizing the full predictive potential of an in-depth functional systems analysis that combines integrated high-throughput data, computational techniques that infer function interactions, and expert biological knowledge manually curated from the literature and public pathway databases.

A Systems Approach to Investigating ESC Self-Renewal

By investigating ESC self-renewal mechanisms at a systems level, we can more effectively unravel how stem cells regulate potency and maintain population sizes appropriate for the rapidly changing developmental needs of embryos. The focus of this project is to apply a proven, statistically principled systems approach for deciphering self-renewal processes in mouse and human ESCs, and provide computational tools for researchers to enable more effective manipulation of iPS cells and cancer stem-like cells to improve human health. I employ a statistically principled machine learning approach to develop and compare predictive
functional relationship networks for mouse and human pluripotent stem cells. Machine learning is a broad field, and there are many different analytical methods with many practical applications, ranging from spam filters to optical character recognition [51]. In general, you can apply machine learning to any problem where a pattern exists that you cannot identify mathematically, and you have data. In this work, I use a form of supervised machine learning, called classification, where there are specific examples to learn from and the goal is to predict membership in one or more classes. There are a variety of methods for classification, such as decision trees, support vector machines, and Bayesian networks, all of which have been applied to a wide range of biological problems [52]. I chose to use simple, yet flexible naïve Bayesian network framework for integrating genomic data to make predictions about functional linkages among protein-coding genes in the context of ESC self-renewal. For biological applications, such as this study, this type of approach is fast, produces results that are easy to interpret, and provides useful details on the reliability of input evidential data. However there are tradeoffs in as well. For example predictive details may be sacrificed for processing speed and the general naïve Bayes framework is based on an independence assumption that doesn’t reflect biology reality. Although there are more sophisticated statistical models, such as tree-augmented networks (TANs) or Gaussian graphical models, which could be applied to this biological problem, such methods are computationally more expensive and have not been shown to yield significantly improved results. It is still an open research problem as to whether such methods can outperform simple relevance networks in terms of predictive accuracy [51, 53, 54].
Inferring Biological Networks from Diverse Functional Genomic Data

This study complements and extends work done to date by implementing systems-based methodologies and computational inference tools rigorously tested and proven in studies on pathways in yeast (*Saccharomyces cerevisiae*) [55]. Commonly used as a platform for the development of computational methods and high-throughput experimental techniques, yeast is a compact, self-contained, functionally well-understood model system ideal for testing systems-based techniques [46, 54, 55]. My approach is based on a strategy for Bayesian network machine learning and genomic data integration similar to a method originally developed by Myers *et al.* to study functional relationships among genes and proteins in yeast. It can be distilled into a three step process: integrate genomic evidence, infer a functional linkage network, and predict localized pathways based on inferred relationships [56]. In the Myers *et al.* study, the authors integrated diverse high-throughput genomic and proteomic datasets using a Bayesian network, which weighed each evidence type probabilistically based on its accuracy. They used Bayesian integration to produce a network with confidence-weighted edges between each pair of genes. To mine this network, they incorporated an interactive query tool, based on a network prediction algorithm that identifies network components by finding proteins likely to be functionally related to a user-supplied query set of proteins of interest [56].
Figure 1.4. Naïve Bayesian Networks for Genomic Data Integration. A. The graphical structure of a simple naïve Bayesian network. B. Functional Relationship (FR) is a hidden conditional variable, on which all dataset evidence variables are dependent. The edge weight \((e_{ij})\) represents the probability that the proteins \(ij\) are functionally related given the evidence: \((e_{ij}) = P(\text{FR=yes} | D^1_{ij})\).

At the core of this approach is a fixed, naïve Bayesian network (Figure 1.4A) [52, 57]. A Bayesian network is a machine learning tool for organizing pieces of knowledge and encoding statistical dependence relationships among these pieces of knowledge. Such graphical models, in which each circle represents a node and each directed edge represents a dependence relationship, provide a flexible framework for combining, controlling, and testing effects of different types of observed data and prior knowledge as input.

A naïve Bayes network (Bayes net) is a simplified version of a Bayesian network that has edges \(C \rightarrow A_1, \ldots, C \rightarrow A_n\) where \(C\) is a class variable and \(A_1, \ldots, A_n\) are attributes. In this simple structure, all child nodes are dependent on the parent and independent of each other [57]. This
type of graphical device may be used to organize statistical information and generate probabilistic models of biological functional relationship (FR) networks, which are typically rendered as dense, complex graphs that represent molecular elements as nodes and predicted functional linkages between nodes as undirected edges [45, 52].

The fixed naïve structure employed by Myers et al. modeled each piece of observed genomic evidence within a dataset as an attribute conditionally dependent on the presence or absence of an edge between two proteins in the predicted FR network, and conditionally independent of all other observed evidence data (Figure 1.4B). Genomic data sets were used to compute similarity scores between all pairs of proteins based on prior knowledge of interactions, based on a gold standard derived from GO biological process annotations. Conditional probabilities for each dataset were estimated by counting the frequency of occurrence of positive and negative examples of different association scores. These learned conditional probabilities were then used to compute the posterior probability of a functional relationship between pairs of proteins for which there were no known interactions [56].

**Applying Bayesian Network Machine Learning Approaches to Mammalian Systems**

Using Bayes net methodologies originally developed for yeast [56, 58], several recent studies have performed mammalian protein function prediction by effectively treating mammals as giant, single-celled organisms [59-67]. These groups leveraged functional annotations provided by public resources, such as the GO and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [48, 49], to automatically generate training sets of proteins co-annotated to selected pathways or processes, and they integrated data from a broad range of tissues, cells, and developmental stages. Results of these pioneering efforts have shown that it is possible to predict protein function accurately using this type of generalized approach [60, 61, 63, 68].
However, such methods are not appropriate for predicting cell-type-specific protein function in mammalian systems, in which the same proteins may play very different roles depending on cellular context and developmental stage. This is largely because, historically, if a gene product performs a function in any cellular or in vitro context, it is annotated to that function. Thus, it is unlikely that a simple, co-annotation-based gold standard is appropriate for making predictions about complex, cell-type-specific mammalian gene functions. For example, as of August 2013, in the Mouse Genome Informatics (MGI) database there were 75 GO term annotations associated with the transcription factor encoded by the Stat3 gene, which is known to be expressed in at least 174 mouse tissue types [69]. In mESCs, Stat3 is known to be a key regulator of self-renewal through the LIF-induced JAK/STAT signaling pathway, but in hepatocytes it is involved in a variety of physiological processes, ranging from liver regeneration to apoptosis and metabolism [69-71]. Stat3 mutants have been associated with abnormal phenotypes ranging from temperature homeostasis to eating behavior to sexual reproduction [69]. In the near future, developing resources, such as the Cell Ontology [72], when combined with the functional knowledge in GO, KEGG, and other curation resources, will support the automatic generation of diverse, cell type-specific training sets for mammalian model organisms. Until these resources are available, the most trustworthy method for assembling a cell-type specific gold standard is manual curation, potentially augmented by text mining tools, such as named entity recognition (NER) software [73].
Bayesian Network Machine Learning for Protein Function Prediction in Pluripotent Stem Cells

This study uses a Bayesian network machine learning approach to explore the molecular foundations of pluripotent stem cell self-renewal and cell fate by controlling the inherent complexity of mammalian system by focusing on a single stem cell system at a defined developmental stage in the context of biological processes known to be active in that cell type. Specifically, I developed species- and cell-type-specific data compendiums using only high-throughput data collected using mESCs and hESCs, and manually curated a gold standard training set from the literature by selecting journal articles focused on ESC self-renewal.

Pluripotent stem cells, in particular, are an ideal model system for Bayesian network machine learning applications. Relatively self-contained and clonal, these cells are one of the closest examples of a single-cell, yeast-like system in mammals. These cells have been well-studied in the literature and extensively annotated in model organism databases, and there is a wealth of high-throughput data for these cells in public repositories. Because differences between mouse and human pluripotent stem cells have not yet been fully characterized, they are ripe for comparative analysis using predictive computational techniques.

Contributions and Project Overview

This work provides a strategy for the study of ESCs using high-throughput data. It describes the algorithms and methods used to develop and compare cell-type specific predictive networks for mouse and human ESCs using Bayesian network data integration and machine learning techniques. The remainder of this dissertation discusses in detail the methods used to develop, test, and analyze predictive networks for pluripotent stem cells, and describes how stem cell researchers may use the resulting tools and resources for discovery analysis and data
visualization. Chapter 2 describes the methodology for creating, testing, and analyzing cell-type-specific predictive networks, using mESCs as an initial case study. It summarizes lessons learned about designing a training gold standard, selecting and preprocessing appropriate data, discusses caveats for using standard machine learning metrics for computational validation, and describes methods and tools for visualization and analysis of information in dense, complex biological networks with millions of predicted functional associations. Chapter 3 provides an overview of the data visualization tool, StemSight Scout, which was developed specifically to provide stem cell biologists immediate access to the massive amounts of biological information about functional associations among protein-coding genes within these predictive networks. It highlights the distinctive features of the StemSight Scout visualization tool and provides a tutorial on how to use it for network analysis. Chapter 4 describes our methodology for developing and analyzing comparative predictive networks for mouse and human ESCs. It specifically discusses how to use a master list of protein-coding gene orthologs and a “species-agnostic” training gold standard with species- and cell-type-specific data to produce networks that can be used to mine and compare functional associations among genes in mouse and human ESCs. Finally, Chapter 5 summarizes and concludes this work.
CHAPTER 2

CELL-TYPE-SPECIFIC NETWORK YIELDS INSIGHTS INTO MOUSE EMBRYONIC STEM CELL SELF-RENEWAL AND CELL FATE

Abstract

Self-renewal, the ability of a stem cell to divide repeatedly while maintaining an undifferentiated state, is a defining characteristic of all stem cells. Here, we clarify the molecular foundations of mouse embryonic stem cell (mESC) self-renewal by applying a proven Bayesian network machine learning approach to integrate high-throughput data for protein function discovery. By focusing on a single stem-cell system, at a specific developmental stage, within the context of well-defined biological processes known to be active in that cell type, we produce a consensus predictive network that reflects biological reality more closely than those made by prior efforts using more generalized, context-independent methods. In addition, we show how machine learning efforts may be misled if the tissue specific role of mammalian proteins is not defined in the training set and circumscribed in the evidential data. For this study, we assembled an extensive compendium of mESC data: ~2.2 million data points, collected from 60 different studies, under 992 conditions. We then integrated these data into a consensus mESC functional relationship network focused on biological processes associated with embryonic stem cell self-renewal and cell fate determination. Computational evaluations, literature validation, and analyses of predicted functional linkages show that our results are highly accurate and biologically relevant.

Our mESC network predicts many novel players involved in self-renewal and serves as the foundation for future pluripotent stem cell studies. This network can be used by stem cell
researchers (at www.StemSight.org) to explore hypotheses about gene function in the context of self-renewal and to prioritize genes of interest for experimental validation.

**Introduction**

Stem cells, uniquely characterized by their ability to self-renew and differentiate, are a promising tool for biomedical research and cell-based therapy. These special cells play pivotal roles in many stages of normal organism development as well as tissue homeostasis and repair [6]. The properties of “stemness” have also been observed in unnaturally stem-like cells, including artificially induced pluripotent stem (iPS) cells, immortalized cell lines, and cancers [17, 34, 74]. A comprehensive, systems-level view of pluripotent cell self-renewal processes will not only advance our knowledge of stem cell biology, but also facilitate the development of safer biomedical applications.

During development, gene expression profiles change continuously as stem cells rapidly proliferate, differentiate, and communicate with each other. Terminally differentiated cells have more stable gene expression profiles that reflect their distinct roles within tissues and organs; the molecular composition of these mature cells differs dramatically depending on cellular function [75]. To manage complexity and minimize confounding factors, most mammalian laboratory experiments are limited to a specific cell type, tissue, or system of interest [44]. However, most bioinformatics and systems biology approaches have not yet addressed cell- and tissue-specific concerns.
Computational Methods for Predicting Protein Function Using High-Throughput Data

Machine learning techniques based on high-throughput data integration have been used to predict protein function in mammals with mixed results [52, 76]. Naïve Bayesian networks (Bayes nets), one form of supervised machine learning, have proven successful for gene function discovery as they provide a statistically principled method to model relationships among proteins based on a solid foundation of biological knowledge [51, 52, 54, 57, 77, 78]. Given a training set of prior knowledge (also known as a gold standard) comprised of protein pairs known to be functionally related (positive training examples) and pairs believed to be unrelated (negative training examples) combined with independent, whole-genome high-throughput datasets (observed evidential data), Bayes nets identify significant patterns in the evidence, assess data reliability, and then predict novel protein relationships based on reliable data [56, 58]. This method is straightforward and facilitates precise control over gold standard and evidential data composition for testing.

Using Bayes net methodologies originally developed for yeast [56, 58], several studies have predicted mammalian protein function by treating mammals as homogenous, single-celled organisms [59-67]. These studies were species-specific and focused on integrating diverse high-throughput data from multiple cell types and tissues. They leveraged functional annotations provided by public resources, such as the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [48, 49], to automatically generate training sets. Results of these pioneering efforts showed that protein function can be predicted accurately using a generalized methodology [60, 61, 63, 68]. However, due to the current state of public annotation resources, these methods are not appropriate for predicting cell-type-specific protein function in mammalian systems. This is because, historically, if a protein performs a function in any cellular or in vitro context, it was annotated to that function. For example, the Mouse Genome
Informatics (MGI) database referenced 68 GO term annotations for Stat3, a protein expressed in more than 174 mouse tissues (as of September 2012) [69]. (Appendix S1 lists symbols, names, and synonyms for all genes mentioned and includes references for abbreviations used in this article.) In mESCs, Stat3 is a regulator of self-renewal through the LIF-induced JAK/STAT signaling pathway; in hepatocytes, it is involved in many physiological processes, from liver regeneration to apoptosis to metabolism [69-71]. Stat3 has also been associated with abnormal temperature homeostasis, eating behavior, sexual reproduction, and other phenotypes [69]. The function of Stat3 is likely to be highly dependent on cofactors, signaling pathways, and other cellular states. Thus, machine learning methods can be misled if the tissue-specific role of mammalian proteins is not defined in the training set and circumscribed in the evidential data.

Here, we demonstrate the utility of predicting cell-type-specific protein function for mESCs and discuss the computational challenges of this task. Specifically, we show that Bayesian network integration methodologies (Figure 2.1) are most useful when applied to a focused biological question, such as a single cell type and biological processes known to be active in that cell.

mESCs are an attractive model system for testing cell-type-specific machine learning techniques because they are relatively homogenous, they have been extensively studied, and diverse high-throughput data collected from mESCs are publicly available. Despite these advantages, stem cell systems are highly complex and pose unique analytical challenges. Cultures often contain heterogenous cell types, from undifferentiated self-renewing ESCs to early developmental endoderm-like cells [79]. Given this inherent complexity, machine learning methods cannot be used to produce molecular models with mechanistic details based on high-throughput data. However, they can provide an “impressionistic” view of molecular interactions and hypothesize novel protein associations for experimental validation [60, 61].
Previous mESC-specific computational studies relied on limited amounts of high-throughput input data, all of which was considered equally reliable. For example, the Integrated Stem Cell Molecular Interaction Database (iScMiD), combined data from 12 different studies (mostly ChIP-Chip) to create a consensus network of ~50K edges [45, 80]. Others have constructed networks to investigate aspects of self-renewal by analyzing a subset of growth conditions, perturbations, and data types [18, 47]. In contrast, we use statistical machine-learning techniques to integrate a much larger, more diverse mESC data compendium (representing work from 60 studies and 6 experimental techniques) and identify novel functional relationships among proteins. By focusing on a single stem-cell system, at a specific developmental stage, within the context of well-defined biological processes, we produce consensus predictive networks with greater biological relevance than those made using generalized, context-independent methods.
Figure 2.1. Naïve Bayesian Networks for Genomic Data Integration. A Bayesian network is a machine learning tool for organizing and encoding statistical dependence relationships among pieces of knowledge. A naïve Bayesian network is a simplified version of a Bayesian network in which all child nodes are dependent on the parent and independent of each other. This type of graphical device may be used to combine different types of evidential data and prior knowledge to generate probabilistic models of biological functional relationship networks. In our naïve Bayes net structure, the functional relationship between the pair of proteins $i$ and $j$ ($FR_{ij}$) is a hidden conditional variable (indicating the unknown or hidden probability that these two gene products are functionally associated), on which all dataset evidence variables are dependent, and represents the discretized, observed similarity score in dataset $k$ for proteins $i$ and $j$. The edge weight ($e_{ij}$) represents the probability that the proteins $ij$ are functionally related given the evidence observed in different high-throughput datasets. Strong evidence of a functional relationship between protein pairs, measured by edge weight, indicates that the proteins behave in a similar way given observed patterns in the high-throughput data. The specific nature of that relationship can be deduced by evaluating the type of datasets that contribute to that edge weight, followed by experimental validation.
**Results**

We used a naïve Bayesian network methodology (Figure 2.2) to create a cell-type-specific predictive biological network of protein-coding genes in the context of self-renewal and closely related processes (e.g. pluripotency and cell fate determination) in mESCs. For our training set, we manually curated a positive reference of 2056 pair-wise gene relationships (with a prior of 1) among 354 genes associated with mESC self-renewal or annotated to signaling pathways involved in early embryonic development (Supplemental Table P.1), based on information extracted from 98 recent journal articles (Supplemental Table P.2). We automatically generated a negative reference of 20,560 protein gene pairs (with a prior of 0) not documented to be associated mESC self-renewal. We joined these references together to produce a mESC self-renewal gold standard with a class distribution of 1:10 (positive:negative) that was used to train the Bayes net. For evidential data, we assembled a compendium of high-throughput mESC data, representing 60 independent research studies, including all mouse data used in prior mESC-focused computational efforts (Table 2.1; Supplemental Table P.3). This mESC data compendium consisted of ~2.2 million data points, collected under 992 conditions, using 6 different high-throughput experimental techniques, and encompassing more than 6 billion gene-pair measurements. We used the trained Bayes net to make posterior predictions of functional relationships among 21,291 protein-coding mouse genes based on patterns observed in the integrated evidential data.

**mESC-Specific Network Predicts Novel Self-Renewal Proteins**

The resulting undirected, predictive mESC network of ~226 million gene pairs had 582,789 high-confidence edges with a posterior inference score of 0.9 or higher involving 8980 genes that were predicted to be strongly associated with self-renewal and cell fate in the
context of mESCs. We identified 56 potential hub genes with a scaled degree of 0.55 or higher, 59% of which were novel players not included the positive gold standard. Computational evaluations showed the network achieved 90 percent precision at 10 percent recall, and had an Area Under the Receiver Operator Characteristic (ROC) curve (AUC) of 0.75 (μ=0.7402, SD=0.01317), which is significantly better than random (p-value=2.685E-10) and is competitive with prior mammalian Bayes net efforts (Figure 2.3A) [60, 62]. Standard machine learning metrics and cross validation revealed some evidence of overfitting (i.e. tailoring a solution so tightly to the training data that the Bayes net does not learn to generalize the trend and recognize new examples) (Figure 2.3B). Regularization and bootstrap aggregation minimized overfitting at the cost of reducing the AUC to 0.72 (Figure 2.3C, Supplemental Figure B.1A) [51, 52, 62, 81, 82]. Top ranked, high confidence edges in this network were supported by a diversity of high-throughput data, but predominantly by Protein-DNA binding data similarity profiles (Table 2.1, Supplemental Figure B.3).
**Figure 2.2.** Cell-Type-Specific Data Integration and Machine-Learning Methodology. Our approach is designed to generate reliable and relevant predictive biological networks using high-throughput data limited to a specific cell type and a training gold standard focused on biological processes active in that cell type. This process can be distilled into four basic steps: 1. Collect and standardize cell-type specific data from studies using diverse high-throughput experimental techniques, including microarray gene expression, chromatin immunoprecipitation (ChIP) on chip (ChIP-Chip), ChIP followed by high-throughput-sequencing (ChIP-Seq), affinity purification followed by mass spectrometry (AP-MS), whole-genome small interfering RNA (siRNA) screens, and phylogenetic sequence similarity. For this case study, we focused on mouse embryonic stem cell (mESC) data. 2. Curate a process-specific gold standard training set to provide a baseline for assessing data reliability and significance for related biological processes known to be active in the cell type of interest. Our gold standard training set consists of experimentally validated pair-wise associations between genes and proteins known to be involved in mESC self-renewal, pluripotency, and cell fate determination. 3. Iteratively test and validate networks. A. Use a naïve Bayesian network classifier to perform inference and predict novel gene and protein relationships. Our network predicts pairwise functional associations that influence mESC self-renewal and early developmental processes. B. Validate the accuracy of predicted functional relationships using standard machine learning performance metrics, cross validation, and bootstrapping, followed by evaluation of biological content. Our protocol for assessing networks ensures our results are highly reliable and relevant to mESC self-renewal. 4. Provide community access to analyses and tools. Through StemSight.org, we provide access to network analyses and visualization tools to enable users to further explore networks centered on their genes of interest.
1. Collect and Standardize Cell-Type Specific Evidential Data
   - High-Throughput Data
   - Extensive experimental data from diverse sources
   - Data Standardization
   - Bayes Net training sets in in standardized formats

2. Curate Process-Specific Gold Standard
   - Bayes Net Training Set
   - Pairwise examples of known functional relationships

3. Iteratively Test and Validate Networks

   A. Generate Bayes Nets and Infer Functional Linkages
      - Learned Bayes Net
      - Supervised machine learning
      - Relationship Inference
      - Context-specific functional relationship networks

   B. Evaluate Results Computationally and Biologically
      - Network Analyses
      - Cross validation and bagging to minimize overfitting
      - Biological Analyses
      - Functional analysis of predictions

4. Provide Community Access to Tools and Network Analyses
   - Network Visualization
     - Views of predicted functional associations
   - Underlying Data
     - Statistical evaluation of supporting datasets
   - StemSight.org
     - Web-based portal to tools and analyses
Table 2.1. Summary of Integrated mESC Genomic Data

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Datasets (Platforms)</th>
<th>Conditions</th>
<th>Gene Pairs</th>
<th>% Supporting Top Edges*</th>
<th>Mean Redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Expression</td>
<td>58 (19)</td>
<td>807</td>
<td>4,843,618,683</td>
<td>45%</td>
<td>41%</td>
</tr>
<tr>
<td>Protein-DNA Interactions</td>
<td>16 (10)</td>
<td>183</td>
<td>914,929,016</td>
<td>55%</td>
<td>58%</td>
</tr>
<tr>
<td>Physical Interactions</td>
<td>1 (1)</td>
<td>1</td>
<td>207</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Phylogenetic Profiles</td>
<td>1 (1)</td>
<td>1</td>
<td>123,284,253</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Whole-Genome RNAi Screens</td>
<td>1(1)</td>
<td>2</td>
<td>131,795,730</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

* Top Ranked Edges | Top 0.01% of Edges

Notes: A total of 77 high-throughput datasets were collected from various public sources to create a compendium of mESC-specific data that included 992 conditions (e.g., columns in a microarray matrix) and ~2.2 million data points (Supplemental Table P.3). These data were standardized and integrated into ~6 billion gene/protein pairs, and used as evidential data to generate a predictive mESC-specific network focused on mESC self-renewal and cell fate. Datasets were weighted based on the amount of shared mutual information contained in each as compared to all evidential datasets used by the Bayes net. A low mean redundancy indicates the dataset is highly unique. As observed in previous Bayesian network data integration efforts, genetic and physical interaction data were the most reliable, but also the least common [11]. We strove to assemble a diverse and comprehensive set of mESC data that would provide the most coverage and be highly informative. Protein-DNA Interaction data included chromatin immunoprecipitation (ChIP) followed by microarray hybridization (ChIP-Chip) and ChIP followed by high-throughput RNA sequencing (ChIP-Seq). Top ranked edges were the 639 edges with a rank order of 1 and an inferred edge weight ≥ 0.9999 (Supplemental Figure B.3A, Supplemental Table P.11); the top 0.01% of the network consists of the 22,664 edges with an inferred edge weight ≥ 0.9966 (Supplemental Figure B.3B, dataset contributions to top 0.01% edges available at StemSight.org/stemdata.html).
Figure 2.3. Network Performance Evaluations. A. Computational assessment of network performance using standard machine learning metrics showed that precision at 10% recall was 90%, and 60% at 25% recall, before and after regularization and out of bag averaging to correct for overfitting to noise. The area under the Receiver Operating Characteristic (ROC) curve (AUC) for the mESC network was 0.7479; after regularization and out of bag averaging, the AUC was 0.7165. B. We conducted 4-fold network cross validations by removing 25% of edges in the gold standard (4-fold Gold Standard). ROC curves showed a small amount of overfitting, most apparent in cross validations for which we removed 25% of genes (rather than edges) from the network training set (Supplemental Figure B.1). C. We conducted 20 bootstrap runs, using a 70-30 split (training to test) of the gold standard answer file, and performed out-of-bag averaging to produce a single network. The relatively flat trend of AUC over out-of-bag-averaging runs confirms the minimal amount of overfitting and produced a single network with high confidence inference scores.
To assess biological content and functional relevance of our mESC network, we used functional genomics tools [69, 83] to evaluate GO term enrichment, validate that gene pairs known to significantly influence mESC self-renewal were strongly connected in the probabilistic network, and identify novel genes with strong functional linkages supported by evidential data. We used standard methods for weighted network analysis to investigate network topology, identify major hubs, and search for novel interactors. Our results confirm the roles of genes and proteins known to be involved in early developmental transcriptional regulation and stem cell maintenance, including \textit{Pou5f1} (also known as \textit{Oct4}), \textit{Sox2}, \textit{Nanog}, \textit{Klf4}, \textit{Suz12}, \textit{Phc1}, and \textit{Trim28}, all of which were major hubs in our mESC-specific network. Functional annotation analysis showed the most strongly connected genes in the mESC network were highly enriched for stem-cell-related biological processes, including development, maintenance, and differentiation, as well as transcriptional regulation (Supplemental Table P.4).

To identify the tightly connected “core” of our gold standard training set, we evaluated the distribution of predicted posterior edge values and identified a edge cutoff of 0.25 (Supplemental Figure B.2A). We then calculated a functional correlation score to the 356 genes involved in positive gold standard edges to identify top self-renewal gene hubs. We consider any training set gene with at least one strong connection (≥ 0.25) to another member of the training set as a more reliable member of the gold standard: a “golden” gold standard gene. We used a “guilt by association” metric to measure the strength of functional linkage between a given gene and genes in our “golden” gene set. We refer to this measure as the Self-Renewal Correlation score (SRC; details in \textit{Materials and Methods}), which we use to evaluate the likelihood of novel gene association with self-renewal programs and to reassess the role of genes included the gold standard (Supplemental Figure B2.C). For example, many genes one would expect to have high SRCs are key self-renewal players such as \textit{Pou5f1} (SRC: 1.0000), \textit{Sox2}
(SRC: 0.9505), and Gdf3 (SRC: 0.9419), while others exhibited low SRCs, such as Pdc (SRC: 0.0104) and Smad9 (SRC: 0.0451), and are thus less likely to be involved in self-renewal and closely related early developmental processes.

In addition to the known gold standard genes, we found many novel genes exhibited high correlation to self-renewal proteins based on network connectivity and SRC (Table 2.2). These genes included: Gbx2, Jarid2, Tcea3, Tdgf1, Msh6, Slc3a2, Ifitm1, Tdh, Reep3, Jam2, Rpp25, Trh, Msx2, Zfp428, Tfp2l1, Etv5. This list is enriched for genes known to play a role in cell fate determination and other early developmental processes as well as genes involved in transcriptional regulation and DNA binding. For example, Jarid2 (SRC: 0.9235) is a regulatory subunit of Polycomb Repressive Complex 2 (PRC2), which is involved in repression of genes important for development and cell fate specification [84]. Gbx2 (SRC: 0.9367), a transcription factor linked to stem cell pluripotency and differentiation in developing embryos, is a direct target gene of WNT signaling known to be involved in neural crest induction as well as specification and formation of the neuroectoderm [85, 86]. However, none of these genes were included in our training set, demonstrating the ability of the Bayes net to predict potentially meaningful novel players in this biological context.
### Table 2.2. Candidate Genes for Experimental Validation

<table>
<thead>
<tr>
<th>Novel Gene</th>
<th>SRC (Scaled)</th>
<th>Novel Gene</th>
<th>K (Scaled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gbx2</td>
<td>0.9367</td>
<td>Gbx2</td>
<td>0.6259</td>
</tr>
<tr>
<td>Jarid2</td>
<td>0.9235</td>
<td>Tcea3</td>
<td>0.6399</td>
</tr>
<tr>
<td>Tcea3</td>
<td>0.9234</td>
<td>Msh6</td>
<td>0.6292</td>
</tr>
<tr>
<td>Tdgf1</td>
<td>0.9013</td>
<td>Zfp296</td>
<td>0.6067</td>
</tr>
<tr>
<td>Msh6</td>
<td>0.9080</td>
<td>Socsc2</td>
<td>0.5583</td>
</tr>
<tr>
<td>Slc3a2</td>
<td>0.9085</td>
<td>Slc3a2</td>
<td>0.6124</td>
</tr>
<tr>
<td>Ifitm1</td>
<td>0.9092</td>
<td>Tdh</td>
<td>0.6087</td>
</tr>
<tr>
<td>Tdh</td>
<td>0.9056</td>
<td>Rpp25</td>
<td>0.5973</td>
</tr>
<tr>
<td>Reep3</td>
<td>0.8971</td>
<td>Tdgf1</td>
<td>0.5184</td>
</tr>
<tr>
<td>Jam2</td>
<td>0.8945</td>
<td>Zfp428</td>
<td>0.5927</td>
</tr>
<tr>
<td>Rpp25</td>
<td>0.9047</td>
<td>Rhob</td>
<td>0.5820</td>
</tr>
<tr>
<td>Trh</td>
<td>0.8877</td>
<td>Reep3</td>
<td>0.5844</td>
</tr>
<tr>
<td>Msx2</td>
<td>0.8932</td>
<td>Akap12</td>
<td>0.5771</td>
</tr>
<tr>
<td>Akap12</td>
<td>0.8865</td>
<td>Ifitm1</td>
<td>0.5816</td>
</tr>
<tr>
<td>Cited2</td>
<td>0.8946</td>
<td>Jam2</td>
<td>0.5798</td>
</tr>
<tr>
<td>Etv5</td>
<td>0.8855</td>
<td>Mcl1</td>
<td>0.5731</td>
</tr>
<tr>
<td>Crmp1</td>
<td>0.8754</td>
<td>Nolc1</td>
<td>0.5729</td>
</tr>
<tr>
<td>Mcl1</td>
<td>0.8843</td>
<td>Abcc4</td>
<td>0.5739</td>
</tr>
<tr>
<td>Rhob</td>
<td>0.8964</td>
<td>Dnmt3l</td>
<td>0.5693</td>
</tr>
<tr>
<td>Gjb3</td>
<td>0.8907</td>
<td>Plcg2</td>
<td>0.5671</td>
</tr>
<tr>
<td>Zfp428</td>
<td>0.8858</td>
<td>Upp1</td>
<td>0.5571</td>
</tr>
<tr>
<td>Mkrn1</td>
<td>0.8752</td>
<td>Cited2</td>
<td>0.5735</td>
</tr>
<tr>
<td>Zfp296</td>
<td>0.8590</td>
<td>Anp32a</td>
<td>0.5722</td>
</tr>
<tr>
<td>Anp32a</td>
<td>0.8651</td>
<td>Jarid2</td>
<td>0.5657</td>
</tr>
<tr>
<td>Upp1</td>
<td>0.8589</td>
<td>Anp32a</td>
<td>0.5722</td>
</tr>
<tr>
<td>Dbf4</td>
<td>0.8777</td>
<td>Dbf4</td>
<td>0.5708</td>
</tr>
<tr>
<td>H2afx</td>
<td>0.8526</td>
<td>Trh</td>
<td>0.5689</td>
</tr>
<tr>
<td>Zswim1</td>
<td>0.8713</td>
<td>Lpp</td>
<td>0.5507</td>
</tr>
<tr>
<td>Slc7a7</td>
<td>0.8503</td>
<td>Ina</td>
<td>0.5499</td>
</tr>
<tr>
<td>Slc7a3</td>
<td>0.8675</td>
<td>Tcfcp2l1</td>
<td>0.4887</td>
</tr>
</tbody>
</table>

**Notes:** We ranked genes not included in our mESC gold standard by network topology measures: self-renewal correlation (SRC) and scaled network degree (K). We used network degree to identify hubs and topologically important gene nodes, and SRC scores to discover genes functionally related to mESC self-renewal. Our highest confidence potentially novel self-renewal genes (in bold) ranked high in both gene lists and were not yet annotated to biological processes associated with self-renewal or cell fate determination.
L-threonine dehydrogenase (*Tdh*; *SRC*: 0.9058) is one of the less well-studied genes in our list of high-confidence novel gene candidates for experimental validation that was strongly predicted to be involved in self-renewal, pluripotency and cell fate, and tightly linked to many of our “golden” gold standard genes, including *Pou5f1*, *Sox2*, *Nanog*, *Nr0b1*, and *Rif1* (Figure 2.4A, Supplemental Figure B.5). *Tdh* catabolizes threonine into glycine and acetyl-CoA, which is used by the TCA cycle to generate ATP. While there were no GO annotations for this gene based on experimental data at the time we developed our training set, nor articles about the role of *Tdh* in mESCs at the time we created our gold standard, recently published articles confirmed that mESCs are dependent on threonine catabolism to support accelerated cell cycle kinetics [87, 88]. To learn more about the underlying datasets that support functional linkages between *Tdh* and key self-renewal genes, such as *Pou5f1*, we evaluated Bayes net statistics for edge weight and top supporting datasets (Figure 2.4B, 2.4C). These statistics showed that the functional relationship between *Tdh* and *Pou5f1* was supported by ChIP-Chip binding data from five different studies investigating the regulatory circuitry of mESCs and microarray data from a study analyzing mESC differentiation. *Tdh* connections to other golden gold standard genes were largely supported by the same type of ChIP-Chip data (Supplemental Figure B.5). By drilling down to the most reliable datasets, as determined by our machine learning evaluations, we were able to quickly identify *Tdh* as a potential target of the core regulatory circuitry of mESC self-renewal and pluripotency [17, 89] to manage cell-cycle controls during the rapid growth phase of early embryonic development.
Figure 2.4. Data Visualization for Mining mESC Self-Renewal Gene Predictions.

A. Views of Tdh-centric networks created using our StemSight Scout visualization tool, available at StemSight.org. Adjusting Scout network views to display only edges with inference scores of 0.5 and 0.9997 show that the novel gene Tdh is tightly connected to many well-known self-renewal genes in our training gold standard, including Pou5f1, Sox2, Nanog, Nr0B1, and Phc1.

B. Supporting edge info for the Tdh – Pou5f1 edge. Supporting edge information shows that this edge is supported by several protein-DNA interaction (PDI) assays as well as gene expression datasets from a study investigating mESC cell differentiation in different mESC cell lines. For supporting edge detail between Tdh and other gold standard genes, see Figure S5 or explore the Tdh interactome online at StemSight.org/scout.

C. SPELL for StemSight: Mix muscular embryonic stem cells. From a supporting edge information window, you can drill down to the individual gene expression levels in microarray datasets. This view shows how expression data reveals rank-ordered correlations observed between Tdh and gold standard genes Gbf3, Fbxo15, Nr0b1, Phc1, Pou5f1, and Sox2.
We used the SRC metric to select our top 10 candidate genes for experimental validation: Tcea3 (SRC: 0.9234), Msh6 (SRC: 0.9080), Reep3 (SRC: 0.8971), Jam2 (SRC: 0.8945), Crmp1 (SRC: 0.8754), H2afx (SRC: 0.8526), Nolc1 (SRC: 0.8476), Klf9 (SRC: 0.8616), Creb3 (SRC: 0.8447), and Myst2 (SRC: 0.8432). This list of novel genes predicted to be associated with mESC self-renewal includes several transcription factors and chromatin modifiers; all have high SRCs, but no GO annotations related to early embryonic development processes.

**Appropriate Evidential Data is Critical for Useful Network Predictions**

To further assess the impact of cell-type-specific evidential data on Bayes net predictions, we prepared three additional input data feature sets by varying the amount, diversity, and appropriateness of evidential mouse data. We generated test networks for each of these feature sets, using the same mESC self-renewal training set and evaluation metrics as for our mESC-specific network.

To evaluate network performance using a relatively small amount of inappropriate data (not specific to mESCs), we trained a Bayes net using a minimalist library of 16 datasets, representing ~300 experimental conditions (Supplemental Table P.5). This feature set, composed primarily of non-cell-type-specific data downloaded from molecular interaction databases, was similar to evidential data compendiums used for prior functional relationship network projects using mouse data [60, 61] and a mouse gene function prediction competition [63]. Networks generated using this minimalist input data and our self-renewal training set produced the lowest AUC (0.5931) and exhibited the least evidence of overfitting. There was insufficient data to perform regularization, which is best applied to very large-scale data integration [62]. The resulting network contained no notable network hubs and had only 543 edges (involving 446 genes) with an inferred probability of a functional relationship greater than
0.9. Of these edges, only one involved canonical embryonic stem cell self-renewal factors (Pou5f1 – Nanog; weight = 0.9754). The remaining edges were a random assortment of loose connections between genes annotated to disparate functions. These functionally vague results confirmed that limited evidential data, while potentially useful for more general gene function studies [60, 61], are inappropriate for exploring a context-specific cellular process, such as mESC self-renewal.

As a negative control, we assembled a feature set composed of a large amount of inappropriate data: 656 datasets from a broad range of mouse tissues and cell types, excluding mESCs (Supplemental Table P.6). This feature set was composed largely of microarray data and spanned ~13,500 experimental conditions. To further explore the impact of using a combination of any type of mouse data, we created a feature “superset” based on a sprawling compendium of all available high-throughput mouse data, including data from our negative control, minimalist set, and mESC-specific datasets: a total of 810 datasets representing ~14,500 conditions (Supplemental Table P.7). Both the negative control and superset networks achieved higher AUCs than the mESC network (0.88 and 0.86, respectively), but they also exhibited more dramatic evidence of overfitting (Figure 2.5A). This was not unexpected as the number of features in these test sets far exceeded the number of genes in the training set. Subsequently, the Bayes net was able to find patterns in the noise of the input data that most likely did not reflect real biology, often manifested as over-inflated results. Overfitting in networks generated using the superset of input data was even more apparent when trained on randomly generated, negative control gold standards. These test networks all achieved AUCs in the mid-to-high 0.80s; however, overfitting was largely mitigated by regularization and bootstrap aggregation, which reduced test AUCs back to the expected random levels (~0.5) (Figure 2.5B).
Figure 2.5. Importance of Feature Selection in Bayesian Network Machine Learning.

A. Networks trained using the same mESC gold standard but different feature sets had markedly different evidence of overfitting. We generated networks using a minimalist library of 16 datasets composed largely of non-cell-type-specific data from molecular interaction databases, our mESC-specific compendium composed of 164 datasets restricted to mouse mESC data and a small amount of data not specific to any cell type, a superset compendium composed of all mESC training data plus an additional 646 non-tissue specific mouse microarrays, and a negative control compendium containing all datasets except those with mESC data. Using machine learning metrics, we found that the minimalist network achieved the lowest ROC curve AUCs and had the least amount of overfitting. The mESC-specific network achieved a higher AUC, with minimal overfitting. The superset and negative control networks had the highest AUCs, but also showed extreme overfitting with a difference of greater than 0.1 between training and test set AUCs. Bootstrapping followed by out of bag averaging largely correct for overfitting in networks, however network content varied dramatically.

B. Overfitting in Networks with Randomly Generated Gold Standards. Networks trained on randomly generated gold standards performed better than random according to standard machine learning metrics, but 4-fold cross validation revealed these networks had evidence of overfitting that could be corrected for using regularization and bagging techniques.

C. Evaluating Network Differences using Positive Gold Standard Posteriors. A scatterplot of superset versus mESC-only network positive gold standard posterior edge (those with a prior of 1) illustrates that while there is relatively high correlation (Pearson correlation $r = 0.6592$), there is also a broad range of disparity between the two networks. A scatterplot of negative control versus mESC shows that there is less correlation between the two networks (Pearson correlation $r = 0.2311$), and reveals the subset of the training gold standard supported by non-mESC data.
To evaluate the impact of data compendium size as well as composition, we conducted a series of data compendium tests using the type of gene expression and molecular interaction data included as evidence in all three test networks (Supplemental Figure B.4). AUCs for networks generated using incrementally increasing numbers of randomly selected mESC gene expression data sets plateaued at ~0.65 as the size of the compendium reached ~45 datasets (~600-700 conditions). Test networks generated using the same mESC gold standard and different compendiums composed of 60 non-cell type specific mouse datasets achieved slightly higher AUCs than a mESC test network based on 58 mESC specific gene expression datasets, and showed evidence of roughly the same amount of overfitting. However, biologically, these networks were all quite different. Only the mESC gene expression test network had high-confidence edges with a posterior edge weight of 1 involving genes highly enriched for biological processes associated with stem cell self-renewal and embryonic development. Even so, this mESC gene-expression-based network was not as biologically relevant or reliable as networks generated using more diverse mESC evidential data (Supplemental Figure B.4, Supplemental Table P.8).

**Computational Performance Metrics Do Not Necessarily Measure Biological Relevance**

When evaluated using traditional machine learning metrics, the negative control and superset network computationally performed better than the mESC-only network. Even after regularization and bootstrap aggregation, both of these test networks achieved AUCs of ~0.80, as compared to 0.72 for our mESC-specific network. However, they were very different networks, capturing very different flavors of biological information (Supplemental Table P.9). Top network hubs in the negative control network were enriched for biological processes associated with FGFR signaling, MAPK signaling, regulation of cell proliferation, gene expression,
and transcription. In contrast, the superset network (which included negative control data as well as mESC-specific data) was enriched for many of the same documented self-renewal functional associations found in the mESC-only network, but the signal was “blurred” in comparison. For example, out of a total of 21,291 protein-coding genes, Pou5f1, Nanog, Sox2, and Suz12 emerged as the most highly connected network hubs in the superset network, after which there was a steep drop-off in degree (Table 2.3). These four genes were also the top hubs in our mESC network, but they were even more tightly connected, with a higher mean degree. In fact, while much of the gold standard was similarly supported by both the mESC-specific and superset data collections, (Figure 2.5C; Pearson’s correlation $r = 0.6592$, $r^2 = 0.4345$), there was a broad range of functional disparity between the two networks. Not surprisingly, this difference in functional linkage was even more evident when comparing the negative control to mESC posterior gold standard edges. (Figure 2.5C; Pearson’s correlation $r = 0.2311$, $r^2 = 0.0534$). In general, high-confidence gold standard genes in the negative control and superset networks were involved in signaling pathways known to be active in both adult and embryonic tissues (Supplemental Table P.10). The points of agreement between networks indicate that either 1) some genes are so strongly connected in the context of mESC self-renewal that additional non-mESC-specific datasets in the superset did not obscure the signal or 2) these genes are strongly connected in multiple cellular and process contexts. Thus, the additional inappropriate datasets used as evidence in the test networks tended to include information not specifically related to stem cell self-renewal in mESCs, resulting in less focused, less biologically meaningful networks, despite seemingly improved computational performance metrics.
Table 2.3. Comparison of mESC-Specific and Test Network Connectivity

<table>
<thead>
<tr>
<th>Mean Degree</th>
<th>mESC-Specific</th>
<th>Mmu Superset</th>
<th>Negative Control</th>
<th>Minimalist</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 0.30</td>
<td>0.056%</td>
<td>0.018%</td>
<td>0.005%</td>
<td>0%</td>
</tr>
<tr>
<td>0.20 – 0.30</td>
<td>2.404%</td>
<td>0.032%</td>
<td>0.655%</td>
<td>0%</td>
</tr>
<tr>
<td>0.10 – 0.20</td>
<td>22.12%</td>
<td>3.35%</td>
<td>18.12%</td>
<td>0%</td>
</tr>
<tr>
<td>0.00 – 0.10</td>
<td>75.42%</td>
<td>96.60%</td>
<td>81.22%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Notes: The percentage of strongly connected gene hubs (those with a mean degree greater than 0.2 out of a total of 21,291 protein-coding genes) is markedly higher in the mESC-specific network as compared to the superset or negative control networks. Degree is a measure that reflects the number of genes within the network that are predicted to be functionally linked to a given gene. In these networks, which were trained using a gold standard focused on mESC self-renewal, a higher mean degree indicated that the given gene is more likely to interact with multiple other genes, and tended to be enriched for mESC-specific self-renewal processes. Highly connected genes in the negative control network were predominantly annotated to biological processes related to self-renewal functions that are active in all cell types, such as transcriptional regulation, cell proliferation, as well as developmental processes associated with multiple cell types, such as embryonic morphogenesis.

Network Visualization Reveals Novel Functional Relationships

To make our predictive mESC network readily available to the stem cell research community, we created an interactive, online visualization resource at StemSight.org. A flat file of the predictive mESC network, containing all edges with an inferred probability of functional relationship ≥ 0.2 (18,097,736 edges) is available for download at this site. For those uncomfortable working with large graph files, this network may be explored online using StemSight Scout (StemSight.org/scout). The Scout dynamic visualization interface, implemented using ThinkMap visualization technology, highlights potentially novel self-renewal genes by coloring nodes based on their SRC score and illustrates the weight of predicted interactions by coloring edges based on the inferred posterior probabilities. Documented self-renewal genes and edges, those included in our gold standard training set or other curated sets, are also color-coded, making it easy to visually segregate novel from known. If a displayed edge is in the positive gold standard training set, links are provided to the original articles documenting the
relationship. With Scout, users can search for and download information about interactomes centered around a gene of interest, view predicted subnetworks for sets of self-renewal genes identified by previous computational studies, and “drill down” into the data underlying predictions (Supplemental Figure B.4, Supplemental Table P.11). Visualization of underlying gene expression levels is provided through a mESC-specific instance of the Serial Patterns of Expression Level Locator (SPELL) system [90], which reveals gene expression correlations from the mESC microarray datasets used to train the classifier (Supplemental Table P.3).

Through the resources available at StemSight.org, we facilitate analyses for which a greater understanding of the underlying data can be informative, and effectively extend the shelf-life, accessibility, and usefulness of existing high-throughput stem cell data in the literature. Furthermore, by focusing user attention on the most reliable datasets for their area of biology, as determined by our machine learning evaluations, we provide a framework for gene function discovery in the context of mESC self-renewal.

Discussion

Statistical Analysis of Input Data Enhances Relevance of Biological Networks

The supervised machine learning approach we used minimizes bias by statistically evaluating data relevance, including which experimental designs are most appropriate and which conditions are most informative. Bayes nets assigned a statistical level of confidence for each input dataset; regularization filtered redundant mutual information shared among datasets. Our method enabled us to report both the strength of the relationship between gene pairs (edge weight) and statistics that describe which evidential data contributed to each edge
(Figure 2.6A, B). Using this information, we could “cross validate” predictions made in other studies, such as functional linkages between transcription factors that have been computationally validated as essential for mESC pluripotency [18, 45] (Figure 2.7A). For example, our results predict a strong functional linkage between Suz12 – Sox2 (edge weight: 0.9998), but a weak connection between Suz12 – Myc (edge weight: 0.0007). Closer inspection of these edges reveals that although more than half of the top supporting datasets are the same for each edge, the contribution strength of evidence often differs significantly (Figure 2.6C) because our approach includes degrees of co-expression, binding affinity, etc., that are not considered when using a binary network construction approach. In this way, our weighted network provides a view that is closer to biological reality as few genes function in a binary fashion in any system context.
Figure 2.6. Advantages of a Statistically Principled Approach. A. The iScMiD Core20 subnetwork of transcription factors used as bait in the 12 studies included in the iScMiD integrated mESC database [34,35], recreated as an undirected graph using edges available from the iScMiD website. In the iScMiD network, all edges have equal weight and all high-throughput data is considered equally reliable, hence the authors note there may be many false positives. B. The fully connected clique of mESC network posteriors for the iScMiD Core20 transcription factors predicts connections not shown in iScMiD and reveals potential false positives as not all connections are equally supported by the evidential data. For comparison, we checked underlying data for two edges, highlighted in yellow, one of which is not supported in the iScMiD subnetwork (Suz12 – Sox2), and one which is only weakly supported in our mESC-only network (Suz12 – Myc). C. Contrasting detailed information about underlying data supporting the strong functional linkage between Suz12 – Sox2 (Edge Weight: 0.9998) versus the weak linkage between Suz12 – Myc (Edge Weight: 0.0007) shows that top supporting datasets vary from edge to edge and that the strength of dataset contribution to edge weight may differ significantly. (Highlighted rows are datasets that support both edges.)
Figure 2.7. Comparing Subnetworks of WNT Signaling Pathway Participants. A. WNT Signaling Pathway Subnetwork. A model of the WNT Signaling pathway adapted from the curated KEGG pathway for *M. musculus* (Mmu) includes SRCs for Wnt, Frizzled, and Dishevelled pathway participants, illustrating that not all family members are equally supported by evidential data. Curated pathways, which are cell-type agnostic, cannot capture these differences in connectivity. A corresponding network of mESC posterior edges involved in this view of WNT Signaling (created in Cytoscape) demonstrates the variance in edge weights and SRCs in the signaling cascade. B. The same WNT Signaling subnetwork produced using Mmu superset and negative control posterior edge weights and SRCs captures a different picture of connectivity as compared to the mESC network. Far more WNT signaling activity between different WNT family member ligands and Frizzled receptors is evident in the test subnetworks. This may reflect WNT signaling activity observed in data from both mESCs and other cellular contexts in the Mmu superset of features. The influence of WNT signaling in other cellular contexts is even stronger in the negative control subnetwork.
Tailoring Prior Knowledge to a Single Cell-Type Clarifies Results

For our mESC gold standard, we intentionally included interactions that influence closely related developmental processes, especially when cell fate hinges on stoichiometry or epigenetic regulation of a common set of genes. For example, normal expression levels of the Pou5f1 transcription factor support self-renewal, while aberrant over expression (as little as two-fold increase) can induce premature differentiation into primitive endoderm and mesoderm cells, and loss of Pou5f1 induces differentiation to trophectoderm cells [91, 92]. For developmental signaling pathways, we included direct and indirect connections between all pathway participants, unless cell-type-specific players were indicated in the literature. For example, for the WNT signaling pathway, we included all 16 Wnt ligands, all 9 Frizzled receptors and all 3 Dishevelled signal transducers in our mESC gold standard. For JAK/STAT signaling, the literature was more explicit with respect to ligand-receptor pairs; subsequently, our gold standard includes only references to the IL6 class of cytokines and cytokine receptors that are important for mESC biology (LIF, LIFR, and IL6ST) [93, 94].

Despite our extensive manual curation efforts, the examples in our training set are of variable quality and reliability. Some interactions, such as Pou5f1 – Nanog were referenced and observed multiple times, while others such as Yy1 – Cbx2, were included as training examples based on less well established experimental evidence. Subsequently, some of our training edges may ultimately prove unreliable. This is especially true for examples derived from signaling pathways (such as WNT), where we included all possible ligand-receptor pairs in the absence of other information. Based on our results, only a subset of these possible interactions appears to be important for mESC biology.

Our SRC measure may prove a useful tool for assessing which gene family members are more likely to be active in a signaling pathway within the context of a specific cell type or
developmental stage. To demonstrate this application, we extracted gold standard edges involved in the WNT signaling pathway (adapted from KEGG), and ranked possible Wnt, Frizzled, and Dishevelled participants by SRC (Figure 2.7A). Our mESC WNT pathway highlights which components are strongly supported by mESC data and may help identify interactors required to activate canonical and non-canonical WNT signaling cascades that influence mESC self-renewal, pluripotency, and cell fate [95-97]. For example, we observed high SRCs for Wnt11, Wnt5a, Wnt4, and Fzd5, which have been shown to work together in a context-dependent manner to activate canonical WNT signaling [98]. In addition, our network captures relationships among Wnt family members known to work in concert to mediate signaling activity. Wnt5a has been shown to compete with Wnt3a for Fzd2 receptor binding sites [95], and this Wnt5a – Wnt3a – Fzd2 triad is strongly supported by our posterior weights.

Comparison of our mESC-specific-network to the negative control and superset networks further illustrates how predicted functional linkages may be used differentially to identify specific ligand-receptor pairs active in mESCs signaling pathways. While our mESC-specific network predicts only a few specific edges between pathway participants (Figure 2.7A), in our negative control and superset networks, all 16 Wnt ligand genes are almost equally correlated with known self-renewal genes, and all but 3 Wnts (Wnt2, Wnt16, and Wnt10b) were strongly linked with most Frizzled receptors (with the exception of Fzd5) (Figure 2.7B, Supplemental Table P.10). These subnetwork views show that in the superset and negative control networks there was evidence of general WNT signaling activity linking most WNT ligands and receptors in some cellular context. Because all WNT ligand-receptor pairs were documented in our training set, the superset and negative control networks were better able to capture all of WNT signaling in mouse, whereas our mESC network results are more specific to mESC self-renewal.
Our results support the importance of cell-type-specific data integration and manually curated gold standards for Bayes net machine learning techniques, and illustrate how networks can be used to create biological-process focused predictive networks. With a better understanding of the tradeoffs involved with gold standard composition, particularly in terms of traditional machine learning metrics [51, 99], one can develop different, yet complementary training sets to explore different facets of biological relationships in the context of a given cell type. This approach may prove particularly useful for predicting functional linkages among families of genes involved in signaling pathways active in many cells and during multiple developmental stages, and for which experimentally validated knowledge of the role of specific pathway participants is sparse.

**Preparation of Evidential High-Throughput Mouse Embryonic Stem Cell Dataset Compendium**

We preprocessed, normalized, and standardized a comprehensive set of mESC input data from high-throughput experiments using microarrays, ChIP-Chip, ChIP-Seq, affinity purification followed by mass spectrometry (AP-MS), and whole genome small interfering RNA (siRNA) screens, plus molecular interaction and phylogenetic data not specific to any cell type. Collectively, this data represents 992 conditions and 2,258,468 data points. A complete list of data sources used is provided in Table S3. This data was mapped to MGI Gene IDs, and preprocessed into ~6 billion pairwise values used as features for classification. Through this process, each protein-coding gene pair (in each dataset) was assigned a similarity score, based on Euclidean or Pearson correlation distance measures between genes (see Equations 1 and 4). Pearson correlations were normalized using Fisher’s Z-transform, shifted by the mean, and divided by the dataset standard deviation to yield a collection of pairwise similarity scores with
an approximately normal distribution ~N(0,1). Values were binned into discrete ranges for use as classification features in our Bayes net integration (details as follow).

**Microarray Expression Data.** Raw mESC microarray expression data files were downloaded from the Gene Expression Omnibus (GEO) [105]. Microarray datasets available in Affymetrix CEL format were normalized using the Robust Multichip Average (RMA) function in Bioconductor R/affy package (version 2.5, R version 2.10.1). Brainarray ENTREZG custom chip definition files (CDFs), which reflect the most recent gene and probe sequences, were used to map probes to genes (version 12.1.0) [62, 106, 107]. To standardize microarray data downloaded from public databases, we followed this protocol: 1) Impute missing values and remove probes with few values (probes were required to be present in at least 70% of conditions to be retained) using the KNN-impute algorithm, 2) Map microarray Probe IDs to systematic MGI gene IDs, 3) Average together consistent probe values using a maximum likelihood approach, and 4) Perform numeric clean up and consolidation (as previously described) [62, 90, 108-110]. The resulting standardized datasets were then converted to a PreClustered (PCL) format and distilled into a set of pairwise similarity scores using Pearson correlation followed by Fisher’s z transformation to measure the strength of the linear relationship between gene expression values for all possible gene pairs in the study (Equations 2.1, 2.2).

\[
r = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \frac{\sigma_x \sigma_y}{\sigma_x \sigma_y} \\
\]

(2.1)

\[
z = \frac{1}{2} \ln \frac{1+r}{1-r} \\
\]

(2.2)

Where \( r \) is the Pearson correlation coefficient calculated from the microarray profiles, and \( z \) is the Fisher Z-transformed correlation.
Chromatin Immunoprecipitation (ChIP) followed by Microarray (ChIP-Chip) Data. Raw high-throughput ChIP data was obtained from online supplemental materials and from contributing author websites. We organized data into a consistent ChIP data matrix format, mapping bait and target gene IDs to systematic MGI gene IDs. Data was processed into a pairwise format in two ways: first, we generated separate data files for each transcription factor used in the study (e.g. input pairs connecting each transcription factor with its putative targets as inputs or features); and second, we generated transcription-factor-binding similarity profiles between all gene pairs (e.g. values determined by the number of transcription factors shared by each gene pair). These similarity profiles were created by calculating dot products between vectors of individual transcription factor binding scores for each gene pair in the study (Equation 2.3).

\[ D_{a,b} = \sum_{i=1}^{n} (a_i b_i) = (a_1 b_1) + (a_2 b_2) + \ldots + (a_n b_n) \]  

Where \( D_{a,b} \) is the dot product score for the pair of genes \( a \) and \( b \), \( n \) is the total number of transcription factors interrogated in the study, and \( a_i \) and \( b_i \) are binding scores (often binary) for genes \( a \) and \( b \) and the \( i^{th} \) transcription factor.

ChIP Followed by High-Throughput Sequencing (ChIP-Seq) Data. Transcription factor binding site and gene association scores [89], based on the genomic location of the binding site closest to the transcription start site of expressed genes, were used as raw data and processed in the same manner as ChIP-Chip data.

Whole Genome Small Interfering RNA (siRNA) Screen Data. Raw siRNA data from primary screens represented the percent of differentiating cells upon exposure to siRNA knockdown of mESC self-renewal genes. We organized these values into a matrix format analogous to that
used to preprocess microarray and ChIP data. A Euclidean distance measure was used to distill this data into pairwise similarity scores (Equation 2.4).

\[ L_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]  

(2.4)

Where \( L_2 \) is the Euclidean distance function calculated between mESC differentiation status levels when genes \( x \) and \( y \) are knocked down.

**Cellular Context Matters in Predictive Biological Networks**

Our mESC network is clearly enriched for self-renewal and early developmental processes, and the top, most highly connected hubs (Pou5f1, Nanog, Sox2, and Suz12) are genes experimentally validated to influence mESC self-renewal. In contrast, other efforts utilizing more generalized GO-based gold standards for gene function prediction capture a completely different connectivity picture, where the major hubs are Brca1, Trp53, and Rb1, and the core self-renewal transcriptional regulator Pou5f1 is connected only to Nanog [61]. This is not to say that general co-annotation training sets are not informative, but rather to emphasize that in complex biological systems, context matters. These approaches have demonstrated value for gene function prediction, but they may not be the best choice for exploring more specific functional associations within a defined cellular context.

The disparity between computational performance and network relevance observed in our test networks is most likely because of the composition of our gold standard training set, which included not only direct edges experimentally validated to be involved in mESC self-renewal, but also more generic direct and indirect edges associated with signaling pathways (such as WNT) active in many tissue types, not just mESCs (Figure 2.7B). These more general
relationships were strongly supported by the non-mESC data included as evidence in our negative control and superset test networks, and subsequently credited as correct results in our computational performance evaluations. Conversely, the absence of evidential support for these generic edges in the mESC network (Figure 2.7A) resulted in a lower computational performance score, even though the mESC-specific network provided a more reliable depiction of gene functional relationships within the context mESC biology. The differences among our mESC-specific and test networks demonstrate the tradeoff in using computational performance evaluations that are blind to cell-type-specificity and the challenge of gold standard development for context-specific network prediction, especially when the goal is to not only recapitulate what is known, but also to discover novel biology at the cellular level.

**Conclusions**

We have shown that naïve Bayesian networks trained using a biological-process-specific gold standard and cell-type-specific evidential data can provide useful, testable insights into novel biology. Our results underscore that traditional machine learning performance metrics alone are not sufficient for evaluation of the predictive accuracy of complex biological networks, particularly when the goal is discovery of novel gene/protein interactions within a defined cellular context. High AUCs may reflect how well a network recapitulates what is known, but they are not the best measure of unknown biology, which doesn’t always play by predictable rules [77, 100, 101]. Reassessments of functional relationship network predictions have observed that study biases, annotation biases, data correlation structures, and high levels of noise can easily mislead machine learning approaches and, consequently, impair biological interpretation of results [100, 102, 103]. In this work, we address many of these potential
pitfalls. By manually curating our training sets, we avoid annotation biases in resources, such as GO and KEGG. By restricting evidential data to our cell type of interest, we reduce the impact of multi-functional genes. We perform extensive network regularization to manage data correlations and biases, and do not rely solely on traditional machine learning performance metrics to assess network quality. In this way, we are able to generate predictive biological networks that more closely reflect biological reality than other, more generalized approaches can achieve.

It is vital to assess the biological relevance of predictive networks in terms of the cellular context of interest. Just as no one type of experiment can elucidate all facets of biological pathways and mechanisms, no one network, regardless of its computational performance, will excel at making predictions about gene function in all contexts. As such, biologists should be both wary of and savvy about which computational tools and databases best support their research efforts, and preferably, use a consensus approach involving multiple computational resources. Predictive networks, such as ours, can aid in preliminary analyses by providing a comprehensive view of information otherwise lost in vast repositories of high-throughput data, but they should not be the only reference tool or method used.

In this study, we suggest alternative analytical approaches that can be used to assess novel biological predictions. We demonstrate the importance of investing in manual curation, not just in terms of gold standard creation, but also for evaluating, restricting, and normalizing datasets for integration. We also highlight the limits of gold standard curation in cases where our knowledge is incomplete, and suggest strategies to identify unsupported training edges, such as those in the WNT signaling pathway, and to tease out novel interactions for potential inclusion in future training sets (as determined by SRC scores). We are currently experimentally validating top candidate genes identified through our computational and functional analyses.
Moving forward, it will be important to extend our cell-type-specific approach to additional cellular contexts. Given the complexity of self-renewal processes and the importance of cellular context, a natural extension of this work would be to evaluate, compare, and contrast the underlying molecular foundations of self-renewal in the context of different stem cell types. As high-throughput techniques and additional resources become increasingly more sophisticated and affordable, computational methods will, in turn, become even more biologically informative. Single molecule sequencing, high-throughput proteomics, flow-cytometry-sorted stem cell populations, single cell data collections, the Knockout Mouse Project [104], and the Cell Ontology [72] will all contribute to the increasing quality, breadth, and depth of consistent, developmental stage specific mammalian data. With these evolving high-throughput data, machine-learning methods such as ours will be able to produce more mechanistic predictive models that trace molecular interactions during early development and throughout a stem cell lineage.

**Materials and Methods**

**Collection and Preparation of Training and Evidential Stem Cell Knowledge**

Supervised Bayesian network machine learning requires a consistently integrated collection of diverse high-throughput evidential datasets, coupled with a reliable reference gold standard (prior knowledge) for training and evaluation. To ensure high quality, consistent, and comprehensive system input, we developed a rigorous protocol for gathering and preprocessing input datasets, and carefully documented our methods for developing tailored gold standards.
**Prepare Curated, Tissue-Specific Training Gold Standard**

We developed a comprehensive positive gold standard for mESC genes and gene products involved in self-renewal, curated from a literature library of 98 recent articles related to mESC self-renewal, pluripotency, and cell fate determination (Supplemental Table P.1). This library was supplemented with mouse developmental pathway information from KEGG [49]. A list of all publications referenced in the mESC gold standard is available in Table S2. From this mESC reference, we extracted a list of 2056 “positive pairs” of gene or gene products experimentally validated to be functionally related in the context of mESC fate. To generate a negative gold standard, we developed a list of 21,291 protein-coding genes (Supplemental Table P.12) derived from an MGI Sequence Coordinates report (MGI_Coordinate.rpt for Build 37, downloaded December 23, 2010) by selecting only MGI IDs that fit the following criteria: Marker Type – Gene, mapped to specific NCBI Gene Start and Stop coordinates, with an official symbol not prefaced by GM (for predicted Gene Marker), and for which at least one evidential dataset contained measurements. This gene list was used to randomly generate a list of “negative pairs” (excluding positive pairs) 10 times the size of the positive gene pair list. The resulting gold standard answer file (Supplemental Table P.13), consisting of a total of 22,616 gene pairs, was used as the prior knowledge to train the Bayes net.

**Construct Bayesian Network and Infer Posterior Functional Relationship Scores**

To perform naïve Bayesian network machine learning techniques, we computed the posterior probability of a functional relationship between gold standard gene/protein pairs given all evidential data [55, 60, 62, 110]. We used the Sleipnir library of C++ tools for machine learning over genomic data [110] and the Structural Modeling, Inference, and Learning Engine (SMILE) C++ library, developed at the University of Pittsburgh [111]. Additional procedural
details on using Sleipnir tools for data integration and network inference are provided in the Supplemental Notes (File S6).

**Bayesian Network Training and Inference.** Conditional probability tables (CPTs) for each dataset were learned by counting the observed values in each dataset’s discretized bins for unrelated and related training gene pairs [51, 62]. Once learned, these CPTs were used to infer posterior functional relationship scores between pairs of genes or gene products. The posterior probability that two protein-coding genes participate in a self-renewal related biological process, given existing data, was calculated based on the prior probability of a functional relationship between genes and the conditional probability of observing evidential data given functional relationship status (Equation 2.5) [51, 55].

$$P(FR | E_1, E_2, ..., E_n) = \frac{1}{Z} P(FR) \prod_{i=1}^{n} P(E_i | FR)$$  \hspace{1cm} (2.5)

Where FR is a hidden variable representing whether a gene pair is functionally related, \(P(FR=1)\) is the predicted probability that a pair is functionally related, \(E_i\) represents the evidence score of the gene pair for the \(i^{th}\) dataset, and \(Z\) is a normalization factor.

**Minimization of Network Overfitting**

We performed four-fold cross-validation (on both edges and gene in the gold standard) and leave-one-gene out cross-validation (also called “jack knifing”) experiments to determine classifier performance and generality. To divide the training set into folds, we used two schemes: the first, randomly separating edges into sets, regardless of which genes were involved in those edges; the second, randomly eliminating a quarter of genes in the genome (more specifically the protein-coding gene list) by removing all edges in a training set fold that
contained those genes. For four-fold edge cross validation, we partitioned the gold standard edges into four randomly generated test sets, while preserving the 1:10 class distribution (positive:negative ratio). We trained classifiers on three folds of the gold standard, using the withheld fold as a validation set, repeating this process four times so that all gold standard training edges were used for both training and validation, and each test fold was used for validation once. We conducted four-fold gene cross validation in a similar manner by partitioning the list of 21,291 protein-coding mouse genes into four gene folds, then creating four test sets, each including only edges involving genes within one fold, while corresponding training sets included edges only between genes in the remaining three folds. For leave-one-gene-out cross validation, we removed one well-known self-renewal gene (Lif, Nanog, Pouf51, or Sox2) and all edges involving that gene to create a test set, while all remaining genes were used for training.

To minimize overfitting, we performed Bootstrap aggregation (i.e. bagging) by 1) creating a series of 20 training and test gold standard files, each consisting of a random 70-30% split of the gold standard file, 2) performing bootstrap runs using these gold standard files (the number of bootstrap runs was determined by the point at which the network performance leveled off), and 3) averaging the inference scores for each gene pair across all bootstrapped networks for which the pair was not used as a training example (i.e. “out of bag” averaging) [82].

**Feature Set Selection**

To assess the importance of input data feature set selection, we compiled three additional libraries of integrated high-throughput mouse data: 1) a minimal feature set consisting of 16 datasets composed of non-tissue specific expression data and information downloaded from online genomic data resources similar to feature sets used for prior efforts
(Supplemental Table S5); 2) a negative control set of 646 datasets excluding mESCs and not specific to any tissue type (Supplemental Table P.6); and 3) a superset of features (labeled Mmu – an organism code for Mus musculus) consisting of all the data used to train our mESC-only network plus all data from the minimalist and negative control compendiums (Supplemental Table P.7) [60, 61, 63]. We performed a full set of performance evaluations, cross-validation, bootstrapping, and out of bag averaging on networks produced using each of these feature sets. We used these alternative networks to compare and contrast network topology and underlying biological meaning of inferred functional relationship scores.

Regularization

Bayes nets impose a strict assumption of independence between input data that is likely violated by many of our input datasets. This limitation can be largely mitigated through regularization of parameters to down weight the contribution of datasets with redundant information (Supplemental Figure B.2). Parameter regularization was performed using mutual information between datasets to weight the strength of prior belief for each dataset [62, 81]. Because the same subset of information could be shared many times among tissue- and context-specific datasets, this regularization provided a quantitative estimate of the amount of redundant information contained in each dataset as compared to all other datasets in the compendium. We calculated a heuristic sum of mutual information relative to the Shannon entropy of each dataset [112](Equation 2.6), exponentially decreasing the weight of a dataset as the amount of shared information increased and incorporated these values into the formula for calculating posterior probability (Equation 2.7) as previously described [51, 55].
\[ S_k = 1 + H(D_k)^{-1} \sum_{i \neq k} I(D_i; D_k) \]  \hspace{1cm} (2.6)

\[ P(FR_{i,j} \mid E_1, E_2, \ldots E_n) = \frac{1}{Z} \prod_{k=1}^{n} \left( \frac{\alpha P[D_k = d_k(g_1g_2)] + \alpha^{S_k} - 1}{\alpha + |D_k| \alpha^{S_k-1}} \right) \]  \hspace{1cm} (2.7)

Where \( S_k \) is a heuristic sum of shared information relative to the dataset’s entropy used to weight the strength of prior belief in a uniform distribution for the dataset, \( H \) refers to Shannon entropy, and \( I(D_i; D_j) \) refers to mutual information. Equation 7 is an variation on Equation 5, such that \( P(FR_{i,j} \mid E_1, E_2, \ldots E_n) \) is the predicted probability that there is a functional relationship between genes \( i \) and \( j \) given evidence in datasets 1 through \( n \), \( Z \) is a normalization factor, \( \alpha \) is a pseudocount regularization parameter used to modulate the strength of regularization required as implied by the strength of the prior (higher pseudocount values weaken influence of redundant datasets), and \( D_k \) is the number of bins used to discretize continuous data values in dataset \( K \). A low \( S_k \) indicated the information contained in the dataset is highly unique, while a high score indicated the datasets contained shared (redundant) information. The redundancy score for each mESC dataset used to train the Bayesian classifier is listed in Supplemental Table P.1. We conducted a series of performance tests, evaluating effects of regularization on similarity score distributions in each evidential dataset and classifier performance, and selected an optimal pseudocount value of 70, which best fit our mESC training set. To produce a similar distribution of posterior edge values for the superset (Supplemental Figure B.2B) and negative control, we used a pseudocount value of 10.
**Computationally Test and Validate Results**

We validated the accuracy of predicted functional relationships computationally using standard machine learning metrics and accepted protocols.

**Evaluation Metrics.** To assess network predictive accuracy, we used standard statistical performance measures for binary (true/false) classification tests: Receiver Operating Characteristic (ROC) Curves, Area Under the ROC Curve (AUC), Precision-Recall Curves (PRC), and Area Under the PRC (AUPRC) [51, 52].

A ROC curve is a two-dimensional graph of true positive rate (TPR) versus false positive rate (FPR) (Equations 2.8, 2.9) that illustrates the relative tradeoff between benefits (true positives, TPs) and costs (false positives, FPs). Precision-recall (PR) curves depict the tradeoff between precision, which is a measure of exactness or quality (*i.e.* how many positive claims are correct), and Recall, which is a measure of completeness or quantity (*i.e.* how many positives were claimed of all possible positives) (Equations 2.10, 2.11) [99].

\[
TP \text{ Rate} = \frac{TP}{P} \tag{2.8}
\]

\[
FP \text{ Rate} = \frac{FP}{N} \tag{2.9}
\]

\[
Precision = \frac{TP}{TP + FP} \tag{2.10}
\]

\[
Recall = \frac{TP}{P} \tag{2.11}
\]

To ensure that network inferences were robust and to assess any evidence of overfitting, we performed four-fold gold standard and genome cross validation, leave-out-one cross validation, and bootstrapping (for details, see “Minimization of Network Overfitting”).
Gold Standard Evaluation. To assess the importance of training the Bayes net using a curated, cell-type-specific gold standard, we generated a series of test gold standards consisting of randomly generated negative pairs and positive genes pairs automatically generated from lists of genes associated with GO terms specific to self-renewal (stem cell maintenance - GO:0019827) and independent of stem cell self-renewal (cellular response to insulin stimulus - GO:0032869, regulation of cardiac contraction - GO:0008016). In addition, we created three gold standards of positive and negative gene pairs randomly generated from our list of protein-coding genes. We trained Bayes nets using these alternative gold standards and the same feature set of mESC data, and used the performance metrics described earlier to evaluate results.

Network Topology Analysis. To analyze the network topology and evaluate biological information contained within graph files, we calculated degree \( k \), sum of degrees \( \Sigma k_i \), mean degree \( \bar{k} \), and scaled degree \( K_i \), for each gene in the training set (Equations 2.12-15) [113].

\[
k_i = \sum_{j \neq i} A_{ij}
\]  
\[
k_{\text{max}} = \max(k)
\]
\[
\bar{k} = \frac{k_i}{n}
\]
\[
K_i = \frac{k_i}{k_{\text{max}}}
\]

Where the degree of the \( i^{th} \) node of vector \( k \) \( k_i \) equals the sum of edge weights between node \( i \) and all other nodes in the training set, and Adjacency matrix \( A_{ij} \) quantifies the connection strength from node \( i \) to node \( j \); the mean degree \( \bar{k} \) is the degree \( k_i \) divided by
the total number of nodes $n$ in the training set; and $k_{max}$ is the maximum degree across all $n$ components of vector $k$.

**Functional Correlation Scores.** For functionally directed analyses, we calculated a functional correlation score $S_i$, for a gene $i$ as the average edge weight between gene $i$ and all genes within a functional set of genes $G$ within a network represented by adjacency matrix $A_{ig}$ (Equation 16).

$$S_i = \frac{1}{G} \sum_{g \in G} A_{ig}$$  \hspace{1cm} (2.16)

For our results, we created two sets of these scores. The first used the set of 354 genes in the positive gold standard to calculate functional correlation scores to the positive gold standard. Based on these scores, we observed that only a subset of our gold standard edges were strongly connected to the rest of the gold standard genes in our results. Therefore, we used an quasi-active-learning approach to refine our set to a highly correlated subset of self-renewal genes, those with a gold standard functional correlation score of $0.25$ or higher (the top 52% of gold standard genes ranked by functional correlation score). Using this subset of 189 strongly correlated genes, we calculated and scaled an updated functional correlation score to this “golden” gold standard set of known self-renewal genes. We refer to this measure as the self-renewal correlation (SRC) score.

We used these values to identify major gene hubs within networks, and segregate clusters of genes that shared similar network properties.
Acknowledgments

The authors are grateful to Curtis Huttenhower for his insightful comments and recommendations regarding Bayes net machine learning and data regularization techniques, and to Judith Blake and Carol Bult for their thoughtful reviews of this manuscript. This project was partially funded by The Jackson Laboratory. KGD is supported by a Pharmaceutical Researchers and Manufacturers of America (PhRMA) Foundation Fellowship in Informatics, and has received funding from the University of Maine Graduate School of Biomedical Sciences, and the National Science Foundation Integrated Graduate Education and Research Traineeship (IGERT) program grant 0221625 to conduct this work. MAH is an Ellison Medical Foundation New Scholar in Aging; he is partially supported by National Institutes of Health /National Institute of Arthritis and Musculoskeletal and Skin (NIH/NIAMS) grant R21 AR069981-01 (Hibbs, Ackert-Bicknell), and National Institutes of Health (NIH) Grant P50 GM076468-06 (Churchill).

Additional Resources

- Database for Annotation, Visualization, and Integrated Discovery (DAVID), http://david.abcc.ncifcrf.gov/
- Graph Algorithms Pipeline for Pathway Analysis (GrAPPA), http://grappa.eecs.utk.edu/
- Mouse Genome Informatics, http://www.informatics.jax.org
- Structural Modeling, Inference, and Learning Engine (SMILE), http://genie.sis.pitt.edu/
- Thinkmap Visualization Technology, http://thinkmap.com
- Weighted Correlation Network Analysis (WGCNA), http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/
CHAPTER 3

STEMSIGHT SCOUT: DYNAMIC DATA VISUALIZATION FOR PLURIPOTENT STEM CELL DATA ANALYSIS AND GENE FUNCTION DISCOVERY

Abstract

Systems biologists must create clear, meaningful, responsive visualizations of complex network data to promote biological discovery without overwhelming biologist end-users with extraneous information. To provide community access to our pluripotent stem cell networks, which may contain more than 20 million edges, representing predicted functional linkages among ~20 thousand protein-coding genes, we created a powerful new network visualization resource (StemSight Scout) based on Thinkmap® visualization technology. StemSight Scout provides an informative and dynamic visualization platform to explore gene neighborhoods within our dense predictive networks, access evidential data underlying network predictions, and link out to external resources for further analysis. StemSight Scout is available online at www.stemsight.org/scout.

Introduction

Visual analysis of dense graphs is vital for extracting biological meaning from complex genomic relationships inferred in predictive biological networks created from large-scale data integration. Several existing online tools provide network views centered on user-specified query genes, including Cytoscape Web [114], Graphle [115], and Medusa [116]. However, visually navigating through online networks generated with these tools can prove challenging
for users, due to overwhelming amounts of information, limited usability features, and the lack of dynamic behavior for real-time interactive exploration. Here, we present StemSight Scout: a new network visualization system, specifically designed to make information in large, complex biological networks readily accessible to biologist end users. StemSight Scout provides dynamic “point-and-click” query refinement capabilities that enable users to expand/contract edges around individual genes of interest. It increases biological usability and interpretability of dense network graphs through search histories, information overlays, visualizations of related datasets, and progressive search capabilities. StemSight Scout provides an easy-to-use interface that facilitates exploration of our dense, cell-type-specific predictive networks for pluripotent stem cells in the context of self-renewal and cell fate determination [1].

StemSight Scout facilitates network analyses for which a greater understanding of the underlying data can be informative, and effectively extends the shelf-life, accessibility, and usefulness of existing high-throughput stem cell data in the literature. By focusing user attention on the most reliable datasets, as determined by our machine learning evaluations, we provide a framework for gene function discovery in the context of a specific cell type and biological processes active in that cell type.

**Implementation**

The StemSight Scout dynamic visualization interface was implemented using the Java-based ThinkMap Software Development Kit (SDK, version 2.9.1, [www.thinkmap.com](http://www.thinkmap.com)) with a J2EE Web application applet plus client-server configuration. On the client-side, Scout is launched in a Java applet that will run on Internet browsers that support the latest version of the Java Runtime Environment (JRE), including recent versions of Firefox, Safari, and Chrome for
Windows and OS X operating systems. We adapted the Thinkmap “spider map” configuration to display our dense, undirected, weighted network data as nodes and edges. All network features (nodes, edges, gene lists, etc.) are interactive and respond to mouse hovers and clicks, providing users immediate access to supporting information contained in our database that is presented on additional web pages. (To view these web pages, browser settings must be set to allow pop-up windows for the Stemsight.org/scout URL.) This enables dynamic exploration of pairwise gene functional relationship predictions, correlations to known self-renewal proteins (called SRC scores), evidential datasets supporting predicted edges, and external online resources for annotations, including related journal articles in PubMed and original data sources [1].

**Available Features**

StemSight Scout provides a flexible, visually informative interface that enables users to explore fully connected, weighted StemSight networks of genes predicted to be associated in the context of pluripotent stem cell fate. These networks include a mouse embryonic stem cell (ESC) network containing more than 18 million significant edges (with an inferred probability of functional relationship of 0.2 or higher) among ~21K protein-coding mouse genes, and comparative mouse and human ESC networks with ~16 million edges among ~17K protein-coding gene orthologs with one-to-one mapping between mouse and human systematic IDs. Using a color-blind friendly palette, Scout highlights potentially novel self-renewal genes by coloring nodes based on their SRC score and illustrates the strength of predicted interactions by coloring edges by their inferred posterior probabilities. Documented edges between genes included in our positive gold standard are also color-coded, making it easy to visually segregate novel predictions from prior knowledge (Figure 3.1A).
To further characterize the biological meaning of functional associations quantified as edge weights, Scout enables users to “drill down” into data underlying predictions to review which types of data and experimental conditions were used in the top supporting datasets. Clicking on any edge opens a new browser window containing detailed information about the edge, including the inferred edge weight and a list of the top datasets supporting the prediction, with links to source datasets and publications (Figure 3.2A). If the edge is documented in the positive gold standard training set, links to PubMed point the user to supporting journal articles. For our mouse networks, visualization of underlying gene expression levels is provided through a mESC-specific instance of the Serial Patterns of Expression Level Locator (SPELL) system [90] (Figure 3.1B).

With Scout, users can search for and download information about interactomes centered around a gene of interest, retrieve network predictions for genes involved in curated developmental signaling pathways [49], or view subnetworks for curated sets of important regulatory genes identified by landmark computational and laboratory studies, including an extended pluripotency interactome from the Orkin Laboratory [47] and an in-silico ESC regulatory network generated by the Lemischka and Ma'ayan Laboratories [117]. Mousing over one of the Curated Sets listed in the menu on the right side of the Scout interface causes any genes in the current network to be highlighted.
Figure 3.1. StemSight Scout User Interface. StemSight Scout has many tools and display options for producing customized user-directed networks. **A.** The StemSight Scout applet window consists of a query toolbar, a network display pane and scrollable network information panes. The default network displayed is always centered on Pou5f1. To create a new network view, click on any gene in the network view, or enter a gene name in the gene query search box and press Enter or click on the Fetch button. Scout searches on official gene symbol, systematic IDs, synonyms, or partial names with a * wildcard. If more than one gene fits the search criteria, a scrollable window listing all associated genes will appear below the search box. At the top of the network display panel is a CONTROLS window shade that reveals query refinement tools to adjust the maximum number of edges and minimum edge weight (inference score) to display. Press the SHOW CONTROLS button to access these tools. HIDE CONTROLS rolls the shade up. Within the CONTROLS window are buttons for accessing tab-delimited text file lists of Genes & SRCs and Edges & Inference Scores for the current network display, as well as a toggle button for alternating between 3D and 2D views. Mouse over any gene node to highlight it and its direct edges, then right- or control-click on the gene node to access the Node Display Options menu. From this menu, users can adjust node labels and highlighting, freeze or hide nodes, expand or collapse node edges, or access gene detail pages. Three network information panes provide details on your query history, genes in the network view, and curated sets that include at least one gene displayed. **B.** The Edge Detail page provides additional information about the weighted edge between any two genes, including gene IDs with links to gene detail pages, the edge inference score, and a list of the top datasets that support the predicted relationship. For mESC networks, this page includes a link to StemSight SPELL. Click on any edge in the network display to retrieve this page relationship. (Note that depending on your browser settings, the Edge Detail page may pop-up directly behind the Scout Applet window.)
Scout has many distinctive and robust usability features. It offers exhaustive options for gene search capabilities that enable users to “fetch” network views centered on a specific gene by entering the official gene symbol, systematic ID, synonyms, or partial names with a * wildcard. In addition, Scout provides several capabilities for immediate network query refinement. Clicking on any gene node in the network view (or on the corresponding gene ID in the Gene Symbols SRC list) will prompt Scout to re-center the network around the newly selected gene and update the network view with new edge information within seconds. Once all the network data is retrieved, the view “freezes” into a temporarily static view that can be further manipulated by the user. Users can click and drag gene nodes to move them around the network display pane to emphasize specific associations between specific genes or groups of genes. Gene nodes can also be hidden to simplify dense network views, enabling users to create network views that focus only on associations between selected genes. To visualize multi-centered gene networks, users can expand or collapse edges for any displayed gene node within a network view to expand or collapse edges.

In the CONTROLS menu, sliders are provided for adjusting the number of edges (2 to 60) displayed per node as well as setting the minimum inference score (0.2 to 0.999999) for edges to be displayed in the applet window (Figure 3.1A). Note that for network “hub” genes, such as Pou5f1 and Sox2, that are predicted to be very strongly connected to hundreds of genes within the biological context of ESC self-renewal, there may be more edges involving that gene that can be displayed in a single Scout network view due to memory constraints of the Java applet environment. For these major hub genes, Scout is best used to visualize the dense connectivity of the gene interactome. To view all network edges involving highly connected hubs, users can extract edges of interest from network graph files, which we provide online at www.stemsight.org/stemdata.html. We recommend using a static visualization tool, such as
Cytoscape [118], to visualize topology for extremely dense gene-centered networks for further analysis.

For further analysis of user-directed networks, Scout provides buttons in the CONTROLS menu for retrieving lists of Genes & SCRs and Edges & Inference Scores in tab-delimited text format that can be used along with screen shots of network views. There is also toggle button to alternate between 3D and 2D display options. To help users keep track of all the network views they’ve created, a history of gene queries viewed during each Scout session is provided.

The information-rich resources of the Scout network visualization resource enable biologists uncomfortable working with large data files to easily explore hypotheses about gene function in the context of pluripotent cell self-renewal and prioritize genes for experimental validation. It is a powerful tool that illustrates the interpretive value of dynamic visualization of dense biological network data.

**Tutorial: Using Scout to Explore Predictive Networks**

This section provides a step-by-step tutorial on how to use StemSight Scout dynamic visualization tools to create and customize user-directed gene interactomes. In this tutorial, we will explore genes predicted to be associated with threonine dehydrogenase in mouse and human ESCs [87, 88, 119].

To open the StemSight Scout Applet window, go to www.StemSight.org/Scout. This will take you to the StemSight Scout Splash! page, which must remain open while you use Scout. (Links to this page are also available on the StemSight.org website.) When the Java security warning pop-up window will appears, click on the Don’t Block button. The Scout window will appear, with the default Pou5f1-centered network. Scout networks consist of gene nodes that
are connected by weighted edges. A key at the bottom of the Network Display pane indicates the type of cell and network context for the network and provides color gradients used to indicate the strength of self-renewal correlation (SRC) for each node, and the strength of the predicted relationship between gene nodes for each edge. The gene the network is centered on is larger in size. It also appears as the highlighted gene listed in the History pane.

Click and drag the mouse anywhere in the white space around the network to reposition it within the Network Display pane. Mouse over genes in the network display or Gene Symbols|SRC pane to highlight individual gene nodes and edges associated with that gene. To reposition a gene on the network display pane, click and drag the node to the desired location and release the mouse. Mouse over edges to highlight individual edges between any two genes. Click on any edge in the network display to retrieve the Edge Detail page. Note that you can open multiple Edge Detail pages during a Scout session. To minimize desktop clutter, we recommend closing these pop-up pages when you’ve finished reviewing edge information.

Creating and Refining Gene-Centered Networks

To create a new network centered on Threonine Dehydrogenase, enter “Tdh” in the Gene Query search box and press the Fetch button. The search box will turn yellow while it loads network gene and edge information. The interactome has completed loading when the search box turns white and the network freezes in place. Note that Tdh now appears as the last gene listed in the History pane (Figure 3.2A). Click and drag the bottom of the query window up to provide more network display space. Try entering alternative names or IDs for Tdh, such as “Threonine”, “Td*”, or “MGI:1926231”.

The default network display is set to show a maximum of 20 edges per node with a minimum edge inference score of 0.2. To modify these settings, click on SHOW CONTROLS to
access Query Refinement tools. Move the Maximum edges slider to the right to display more edges or to the left to display fewer edges. Similarly, move the Minimum inference score slider to the right to prune the network view to display only genes with higher inference scores (stronger predicted connections), or to the left to show edges with lower inference scores (weaker predicted connections). Adjust the Maximum edges to 10, and the Minimum inference score to 0.9998 (Figure 3.2B). Until modified again, these settings will stay in effect for all future networks viewed during the Scout session. Press HIDE CONTROLS to roll the window shade up.

To hide specific genes in a network view without using query refinement settings, right- or control-click on the gene node and select your preferences in the Node Display Options pop-up menu. For example, to hide B3gnt7 in the Tdh-centered network, right-click on the B3gnt7 node and select Hide Options, Hide selected node. B3gnt7 will be hidden from the current network view. To restore B3gnt7, right-click on any other node and select Hide Options, Unhide all nodes. There are several other display options in this menu that can be used to further refine network views. For example, to focus on network topography, right-click on any node and select Highlight Options, Remove most node labels. This will hide all gene labels except for the center node. In this view, official gene symbols for subsidiary gene nodes appear only when you mouse over a node or gene in the Gene Symbols|SRC panel list. To label only selected gene nodes, right-click on the node and select Highlight Options, Label node. This enables you to focus attention on a few key genes in a network view. To restore all gene symbol labels, select Highlight Options, Label all nodes. The best way to learn about them is to try different options and see how they modify the network display.
Creating Multi-Centered Gene Networks

In the literature, Sox2, Pou5f1, and Rif1 have been identified as key regulators of Tdh in mESCs [119]. To create a multi-centered network for Tdh, Pou5f1, Sox2, and Rif1, right-click on Pou5f1 and select Highlight Options, Show node edges. This will prompt Scout to retrieve edges for those genes and increase the size of the Pou5f1 gene node to indicate it has been elevated to a center node. (To downgrade an elevated node, right-click on the node and select Highlight Options, Collapse node. Repeat this step to show the node edges for Sox2 and elevate it to a center gene. Because Rif1 is not displayed in this network view, slowly increase the Maximum edges per node until Rif1 is displayed (it should appear at 16). Right-click on it to show its node edges, then decrease the maximum edges per node back to 10. Although Rif1 is not directly connected to Tdh, Pou5f1, or Sox2 in this network view, it shares many of the same strong connections to important positive gold standard genes Nr0b1, Zic3, Rest, and Fbxo15. Click on the Sox2 node to re-center the network around Sox2. Using such stringent query refinement settings, Tdh does not appear in the Sox2 network view, but Pou5f1 and Rif1 remain center nodes. Notice that Sox2 is included in several curated sets. Mouse over the Gold Standard (Hibbs Lab) list item to highlight genes in this Sox2 network that are in the positive gold standard. In the history pane, click on Tdh to return to the Tdh-centered network. No curated sets are listed because Tdh is not included in any of the standard signaling pathway and pluripotency gene sets built into Scout.

Scout draws network views that position gene nodes based on their relative connection strengths (Figure 3.2B). To further refine the network view to emphasize gene hubs, specific interconnections, or clusters of interesting genes, click and drag on individual nodes to reposition them on the network display panel. Be careful not to click and release too soon or you may accidentally redraw the network centered on the node you clicked on.
Collecting Network Details for Further Analysis and Reports

To document information about a customized Scout network view, use a combination of Scout tools and system utilities. Click on SHOW CONTROLS, and press the Genes & SRC button to open new webpage containing a list of all gene nodes in the displayed network, including the official gene symbol, systematic ID, and SRC score. The Edges & Inf. Score button retrieves a list of gene symbols, IDs, and inference scores for each edge. These lists may be saved as a tab-delimited text file or copied and pasted into a text editor.

To save network images, use screen shot utilities or system commands. For example, on Apple OSX systems, press Shift-Control-Command-4 to turn the mouse cursor to cross-hairs; click and drag the mouse to select the section of the network you to be copied, then paste it into a document or spreadsheet. On Microsoft Windows systems, use the Snipping tool (click on the Windows icon button in the far left corner, enter Snipping Tool in the search box, then click and drag the mouse to copy the network.

The Edge Detail pages contain information that helps characterize the nature of the functional relationship between two genes in a StemSight network. Click on the edge between Tdh and Pou5f1 to view the supporting information about that edge (Figure 3.2C). To archive this edge information, save the html page as a text file or take a screenshot of the webpage.
**Figure 3.2** Using StemSight Scout to Explore Network Predictions for Threonine Dehydrogenase (*Tdh*) in mESCs [119]. **A.** The default view of a network centered on Threonine Dehydrogenase shows a maximum of 20 edges (with an inference score ≥ 0.2) per gene. **B.** Adjusting the control settings to display only 16 edges (with inference scores ≥ 0.9998) per gene and “expanded” edges for *Pou5f1*, *Sox2*, and *Rif1* provides a more visually informative view. **C.** Detailed information on the datasets supporting the edge between *Tdh* and *Pou5f1* shows that both ChIP-Chip and microarray gene expression data contribute to the predicted strength of the relationship between genes.
C StemSight: Pou5f1 – Tdh edge detail

This page shows the information supporting the relationship between Pou5f1 (MGD: 101893; SRC: 0.9969) and Tdh (MGD: 117262; SRC: 0.9396).

Explore the datasets showing correlation between Pou5f1 and Tdh using StemSight SPELL.

Inference Score
The predicted strength of the relationship is 0.999995.

Supporting Datasets
The top datasets that support this predicted relationship are:

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Type</th>
<th>Description</th>
<th>Data Source</th>
<th>PMID</th>
<th>Redundancy</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.595201</td>
<td>PDI</td>
<td>mESC transcriptional regulation TF binding correlations (MG-40H)</td>
<td>mesc et al.</td>
<td>18052474</td>
<td>0.329998</td>
<td>0.010888</td>
</tr>
<tr>
<td>0.453738</td>
<td>GE</td>
<td>R1 ESC differentiation in vitro (MG-40H)</td>
<td>GEO Dataset</td>
<td>GDS677</td>
<td>0.048522</td>
<td>0.010974</td>
</tr>
<tr>
<td>0.395277</td>
<td>PDI</td>
<td>Pou5f1/Nanos regulatory network TF binding correlations</td>
<td>Loh et al.</td>
<td>1618491</td>
<td>0.000991</td>
<td>0.012430</td>
</tr>
<tr>
<td>0.389907</td>
<td>GE</td>
<td>JLI ESC differentiation in vitro (MG-40H)</td>
<td>GEO Dataset</td>
<td>GDS654</td>
<td>0.048522</td>
<td>0.010974</td>
</tr>
<tr>
<td>0.388390</td>
<td>GE</td>
<td>R1 ESC differentiation in vitro (MG-40H)</td>
<td>GEO Dataset</td>
<td>GDS666</td>
<td>0.050032</td>
<td>0.013432</td>
</tr>
<tr>
<td>0.307732</td>
<td>GE</td>
<td>Embryonic stem cells and embryoid bodies (MG-40H)</td>
<td>Not specified</td>
<td>GEO Dataset</td>
<td>0.048522</td>
<td>0.010974</td>
</tr>
<tr>
<td>0.305039</td>
<td>PDI</td>
<td>Myo interaction network TF binding correlations</td>
<td>Kim et al.</td>
<td>18946986</td>
<td>0.193091</td>
<td>0.044335</td>
</tr>
<tr>
<td>0.290269</td>
<td>PDI</td>
<td>Td3 regulatory network TF binding correlations</td>
<td>Cole et al.</td>
<td>183147094</td>
<td>0.072999</td>
<td>0.020765</td>
</tr>
<tr>
<td>0.262753</td>
<td>PDI</td>
<td>mESC pluripotency network TF binding correlations</td>
<td>Kim et al.</td>
<td>18538816</td>
<td>0.009474</td>
<td>0.024863</td>
</tr>
<tr>
<td>0.235328</td>
<td>GE</td>
<td>Nanog and Oct4 knockdown effect on ESCs</td>
<td>GEO Dataset</td>
<td>GDS11421</td>
<td>0.080949</td>
<td>0.025333</td>
</tr>
</tbody>
</table>
**Future Plans**

StemSight Scout is among the first generation of browser-based, dynamic online visualization tools powerful enough to retrieve network information from massive biological data files in real-time, without requiring users to install software on the client. Unfortunately, like all Java-based applications, it is restricted by memory limitations of the Java environment and plagued by Java security issues. To address these issues, we have upgrades StemSight Scout to the latest version of Thinkmap for HTML5. This upgrade (Release 2.0), which was in beta testing at the time of writing and scheduled to be released in September 2013, will enable us to provide touch capabilities and deliver tablet and smart phone applications for StemSight Scout visualization. StemSight Scout Release 2.0 will include a comparative view, which will enable stem cell biologists to directly compare human and mouse ESC networks in a single visualization interface, and significant enhancements to multi-centered gene search capabilities. Data visualization is a rapidly evolving field. We will continue to test and evaluate promising new visualization tools and software environments as they available.
CHAPTER 4

NOVEL INSIGHTS INTO EMBRYONIC STEM CELL SELF-RENEWAL REVEALED THROUGH COMPARATIVE HUMAN AND MOUSE SYSTEMS BIOLOGY NETWORKS

Abstract

Embryonic stem cells (ESCs), characterized by their ability to both self-renew and differentiate into multiple cell lineages, are a powerful model system for biomedical research and developmental biology. Human and mouse ESCs share many defining features, yet have several distinctive aspects, including fundamental differences in the signaling pathways and cell cycle controls that support self-renewal. Here, we explore the molecular basis of human ESC (hESC) self-renewal using a Bayesian network machine learning method to integrate cell-type-specific, high-throughput data for gene function discovery. We integrated and analyzed high-throughput ESC data from 83 human studies (~1.8 million data points collected under 1100 conditions) and 62 mouse studies (~2.4 million data points collected under 1085 conditions) into separate human and mouse predictive networks focused on ESC self-renewal to identify shared and distinct functional relationships among protein-coding gene orthologs. Computational evaluations using standard machine learning metrics show that these networks are highly accurate, and literature validation confirms their biological relevance. Our results reflect the importance of key regulatory genes known to be strongly associated with self-renewal and pluripotency in both species (e.g. POU5F1, SOX2, and NANOG), identify metabolic differences between species (e.g. threonine metabolism), clarify differences between human and mouse ESC developmental signaling pathways (e.g. LIF-activated JAK/STAT in mouse; NODAL/ACTIVIN-A-activated FGF in human), and reveal many novel genes and pathways predicted to be
functionally associated with self-renewal in each species. These interactive networks are available online at www.StemSight.org for stem cell researchers to develop new hypotheses, discover potential mechanisms involving sparsely annotated genes, and prioritize genes of interest for experimental validation.

Introduction

Embryo-derived pluripotent stem cells are a powerful model system for biomedical research and the study of developmental biology. Among the most studied embryo-derived stem cells are human and mouse ESCs (hESCs and mESCs). Since first isolated from human embryos in 1998 [10], hESCs have been of particular interest to the research community as a tool for manipulating cell fate, analyzing characteristics of induced pluripotent stem (iPS) cells and cancer-stem-like cells, and testing potential medical and pharmaceutical applications [120].

All ESCs express common markers of pluripotency, such as POU5F1 (also known as OCT4), SOX2, and NANOG, but there are many known differences between species, including morphological, molecular, and epigenetic characteristics that are reflected in ESCs grown in culture. hESCs are derived from the inner cell mass (ICM) of blastocysts 5-8 days post fertilization and form flat, two-dimensional round colonies with well-defined boundaries, whereas mESCs are isolated from the ICM of blastocysts 3.5-4.5 days post fertilization and form tight, dome-shaped colonies [121]. hESCs respond to cooperative signaling between NODAL/ACTIVIN-A-activated FGF and TGF-β signaling pathways to sustain prolonged self-renewal in vitro [10, 121, 122]. In contrast, mESCs grown in culture require growth factors, LIF and BMP4, to activate JAK/STAT signaling [8, 123, 124]. Self-renewal in mESCs can be boosted through small molecule inhibitors that block differentiation cues from the FGF/ERK signaling
cascade and mimic WNT/β-catenin signaling [29]. Interestingly, the same culture conditions that support self-renewal in mESCs can drive differentiation in hESCs. For example, hESCs exposed to BMP4 and LIF yield extraembryonic phenotypes, and FGF inhibition promotes neuroectoderm commitment [125]. In general, well-defined, standard protocols exist for growing mESCs in culture using cell lines of similar genetic background (predominantly derived from 129S/P/T substrains) [126, 127]. However, hESC cell lines have been derived from genetically distinct embryos in multiple laboratories, each using different media cocktails and protocols to promote self-renewal and the pluripotent state; consequently individual hESC cell lines may respond differently when grown under the same culture conditions [128].

mESCs share epigenetic traits of preimplantation blastocysts and are said to be at the “naïve” or primitive developmental ground state of pluripotency, prior to X-chromosome inactivation and genomic imprinting [127]. These naïve ESCs show no differentiation bias, can self-renew indefinitely in vitro, and can be expanded clonally without compromising the pluripotent state [129]. In addition, mESCs can be genetically modified, then reintroduced into preimplantation embryos to generate high-grade chimeric mice [24, 127]. In contrast, hESCs are said to be “primed” for differentiation and female cells have typically undergone random inactivation of one X chromosome [127, 130], however, there is variability in imprinted genes and other DNA methylation patterns depending on culture conditions and passage number [130, 131]. Intriguingly, hESCs share many molecular and epigenetic characteristics with mouse pluripotent stem cells isolated from the post-implantation epiblast (mEpiSCs) [24, 28, 129]. Although the chimerism assay may not be used with human cells for ethical reasons, studies have shown that lower primate ESCs cannot produce high-grade chimeras when injected into blastocysts, nor can they contribute to the germline [129, 132], indicating that lineage potential
is limited in vivo. While these phenotypic contrasts are clear, the molecular foundations of naïve versus primed ESCs have not yet been fully characterized.

Understanding the differences between pluripotent cell types is of increasing interest and value as we develop new methods to “reprogram” adult cells to an ESC-like state for ongoing research or therapeutic applications. In vitro methods used to reprogram different types of human and mouse adult cells into induced pluripotent stem cells (iPSCs) vary widely, use different combinations of reprogramming factors (e.g. POU5F1, SOX2, KLF4, KLF2, MYC, LIN28A) and experimental conditions, and yield ESC-like cells with different degrees of developmental potential [12, 133, 134]. In mouse iPSCs, the Nanog gene is critical for both blocking differentiation and achieving a naïve pluripotent state [125]. Human iPSCs have also been shown to require NANOG to block differentiation, but, to date, attempts to derive or reprogram hESCs to a naïve pluripotent state have been unsuccessful [129].

Because naïve mESCs and primed hESCs respond to different signaling pathways to sustain and exit the self-renewing state [129], cross-species systems-level analyses of these cell types can reveal molecular details that will assist researchers in assessing the pluripotent state of embryo-derived cells and reprogrammed adult cells, and reveal novel functional homologs that support self-renewal and related early developmental processes. While many systems biology studies have been conducted to explore the molecular basis of self-renewal and pluripotency through gene expression profiles or regulatory networks of transcription factor binding, these efforts have been largely species-specific, restricted to data from a relatively small number of cell lines, limited to a single experimental platform, and/or focused on a single regulatory characteristic influencing cell fate [17, 121, 125, 135, 136]. The goals of this study are to discover potential novel regulators of ESC stem cell renewal in hESCs and perform a cross-
species comparative analysis to enrich our understanding of shared and distinct molecular characteristics of hESCs and mESCs.

We applied a Bayesian network (Bayes net) machine learning approach based on species- and cell-type-specific data integration, which we previously applied to mESCs [1], to produce a consensus network that predicts gene function associations in the context of hESC self-renewal. Bayes nets, a type of supervised machine learning, are particularly useful for gene function discovery as they provide a statistically principled method to model relationships among genes given a solid foundation of biological knowledge [51, 52, 54, 57, 77, 78]. We generated comparative human and mouse predictive networks for ESC self-renewal using a training set that included both ESC self-renewal examples experimentally validated in hESCs, mESCs, or both species. The results of this study provide insights into novel genes involved in hESC self-renewal as well as new comparisons of the molecular characteristics of the most widely-studied model ESC systems.

**Materials and Methods**

**Master Gene List and Training Set Construction**

Using a homology report from Mouse Genome Informatics (MGI) [69], we created a master gene list of 17,342 protein-coding mouse and human gene orthologs with one-to-one homology associations (Supplemental Table P.14) [137]. Our positive training examples (Supplemental Table P.15) consisted of 2077 manually curated pair-wise gene relationships involving 365 genes associated with mouse and/or human ESC self-renewal based on 108 recent journal articles (Supplemental Table P.16), or annotated to early embryonic developmental
signaling pathways by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [49].

Negative/background training examples were generated by randomly selecting 20,770 gene pairs not involving any positive training example pairs (Supplemental Table P.17) and used to create a training set in a 1:10 ratio of positive:negative examples.

**Evidential Dataset Collection**

We assembled a compendium of high-throughput hESC data, representing 83 independent research studies and consisting of ~1.8 million data points from 55 hESC cell lines (though predominantly H1 and H9), collected under 1100 conditions, using 6 different high-throughput data types, and encompassing more than 12 billion gene-pair measurements (Table 4.1; Supplemental Table P.18). For mESCs, we updated our existing compendium of high-throughput mESC data [1] to include work from 62 independent research studies, consisting of ~2.3 million data points from 35 mESC cell lines (all derived from 129S/P/T substrains), spanning 1085 conditions, using 5 different high-throughput data types techniques, and encompassing more than 7 billion gene-pair measurements (Table 1; Supplemental Table P.19).
Table 4.1. Summary of Integrated Data in Human and Mouse ESC Data Compendiums

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Datasets Platforms</th>
<th>Genes Measured</th>
<th>Conditions</th>
<th>Gene Pairs</th>
<th>Redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>hESC Data Compendium</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Expression</td>
<td>72</td>
<td>16</td>
<td>1,206,852</td>
<td>597</td>
<td>10,409,092,802</td>
</tr>
<tr>
<td>Protein-DNA Interactions</td>
<td>28</td>
<td>8</td>
<td>383,784</td>
<td>74</td>
<td>702,255,900</td>
</tr>
<tr>
<td>Protein-Protein Interactions</td>
<td>3</td>
<td>3</td>
<td>13,648</td>
<td>3</td>
<td>154,876</td>
</tr>
<tr>
<td>Epigenetic Markers</td>
<td>9</td>
<td>1</td>
<td>152,946</td>
<td>194</td>
<td>1,299,505,689</td>
</tr>
<tr>
<td>Phylogenetic Profiles</td>
<td>1</td>
<td>1</td>
<td>2391</td>
<td>229</td>
<td>3267</td>
</tr>
<tr>
<td>RNAi Screens</td>
<td>3</td>
<td>1</td>
<td>42,570</td>
<td>3</td>
<td>44,618,430</td>
</tr>
<tr>
<td><strong>mESC Data Compendium</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Expression</td>
<td>61</td>
<td>22</td>
<td>1,028,918</td>
<td>872</td>
<td>5,309,037,873</td>
</tr>
<tr>
<td>Protein-DNA Interactions</td>
<td>102</td>
<td>14</td>
<td>1,311,167</td>
<td>210</td>
<td>1,507,210,159</td>
</tr>
<tr>
<td>Protein-Protein Interactions</td>
<td>1</td>
<td>1</td>
<td>207</td>
<td>1</td>
<td>207</td>
</tr>
<tr>
<td>Phylogenetic Profiles</td>
<td>1</td>
<td>1</td>
<td>15,703</td>
<td>1</td>
<td>123,284,253</td>
</tr>
<tr>
<td>RNAi Screens</td>
<td>1</td>
<td>1</td>
<td>16,891</td>
<td>2</td>
<td>131,795,730</td>
</tr>
</tbody>
</table>

**Notes:** These data were collected from 83 hESC studies and 62 mESC studies, then standardized and integrated into a pair-wise format, and used as evidential data to generate predictive hESC- and mESC-specific networks focused on ESC self-renewal. Datasets were weighted based on the amount of shared mutual information (redundancy) contained in each as compared to all evidential datasets used by the Bayes net. A low mean redundancy indicates the dataset is highly unique. Dataset contributions to the top 0.001% (1503) edges most differential genes pairs in the hESC and mESC networks are available in Supplemental Table P.23.
Construction of Bayes Nets and Inference of Posterior Functional Relationship Scores

We trained separate naïve Bayes nets using the same gene list and training set, but with species-specific ESC data compendiums to predict the probability of functional associations among 150,363,811 protein-coding gene pairs based on patterns observed in the integrated evidential data. Learning was achieved by computing the posterior probability of a functional relationship between training set gene pairs given all evidential data [55, 60, 62, 110] as previously described (Equation 4.1) [1].

\[
P(FR | E_1, E_2, ..., E_n) = \frac{1}{Z} P(FR) \prod_{i=1}^{n} P(E_i | FR)
\]

(4.1)

Where \( FR \) is a binary variable representing whether a gene pair is functionally related, \( P(FR=1) \) is the predicted probability that a pair is functionally related, \( E_i \) represents the evidence score of the gene pair for the \( i^{th} \) dataset, and \( Z \) is a normalization factor.

Naïve Bayes nets impose a strict assumption of independence between evidence data that is likely violated by many of our input datasets, as such, we performed regularization to quantitatively correct for the amount of redundant information contained in each dataset as compared to all other datasets in the compendium (Equations 4.2 and 4.3) as previously described [1].

\[
S_k = 1 + H(D_k)^{-1} \sum_{i \neq k} I(D_i; D_k)
\]

(4.2)

\[
P(FR_{i,j} | E_1, E_2, ..., E_n) = \frac{1}{Z} \prod_{k=1}^{n} \frac{\alpha P[D_k = d_k(g_1, g_2)] + \alpha^{S_k} - 1}{\alpha + |D_k| \alpha^{S_k-1}}
\]

(4.3)
Where $S_k$ is a heuristic sum of shared information relative to the dataset’s entropy used to weight the strength of prior belief in a uniform distribution for the dataset, $H$ refers to Shannon entropy, and $I(D_i;D_j)$ refers to mutual information. This ultimately results in Equation 3 (a variation on Equation 1), such that $P(FR_{ij}|E_1, E_2,...E_n)$ is the predicted probability that there is a functional relationship between genes $i$ and $j$ given evidence in datasets 1 through $n$, $Z$ is a normalization factor, $\alpha$ is a pseudocount regularization parameter used to modulate the strength of regularization required as implied by the strength of the prior (higher pseudocount values weaken influence of redundant datasets), and $D_k$ is the number of bins used to discretize continuous data values in dataset $K$.

A low $S_k$ indicates that the information contained in the dataset is highly unique, while a high score indicates that the datasets contained shared (redundant) information. The redundancy score for each dataset used to train the Bayesian classifier is listed in Supplemental Tables P.18 (hESC) and P.19 (mESC). We conducted performance tests to evaluate effects of regularization on similarity score distributions in each evidential dataset and classifier performance, and selected a pseudocount value of 70 to regularize mESC evidential data. The redundancy score for each hESC dataset is listed in Supplemental Table P.18; we used a value of 100 to regularize these hESC evidential data to achieve a similar posterior probability distribution.

**Computational Performance Assessment**

To assess biological content and functional relevance of our mESC network, we used functional genomics tools [69, 83] to evaluate Gene Ontology (GO) term enrichment [23], validate that gene pairs known to significantly influence mESC self-renewal were strongly connected in the probabilistic network, and identify novel genes with strong functional linkages.
supported by evidential data. We used standard and custom methods for weighted network analysis to investigate network topology, identify major hubs, and search for novel functional interactors [1, 113]. These computational validations, including standard machine learning metrics and cross validation, show that our networks are highly accurate and a powerful tool for comparative analysis (Supplemental Figure B.6).

**Network Topology and Correlation Analyses**

To analyze the network topology and evaluate biological information contained within graph files, we calculated degree \(k\), sum of degrees \(k_i\), mean degree \(k_{\text{mean}}\), and scaled degree \(K_i\), for each gene as previously described [1, 113]. To determine genes strongly associated with self-renewal (rather than the general connectivity measured by degree), we identified a set of “core” self-renewal genes from our positive training set for each species as those genes tightly connected to other training set genes as previously described (173 mESC genes; 111 hESC genes) [1]. Then for every gene in the genome, we calculated the average posterior probability of functional relationship to these “core” self-renewal genes, which we refer to as the self-renewal correlation (SRC) score. We used these SRC values to rank genes from most strongly correlated to self-renewal (ranks near 1) to least correlated to self-renewal (ranks near 17,342). We produced a differential network by subtracting the posterior edge weight for each gene pair (mESC minus hESC probability) to identify pairs strongly associated in one network, but not the other. Spearman correlation coefficients were used to determine the conservation of edge weights between networks.

**Supplemental Materials and Methods**

Further details of Materials and Methods are provided in Appendix D.
Figure 4.1. Bayes Net Machine Learning Methodology for Cell-Type-Specific Comparative Networks. To produce reliable, biologically relevant comparative predictive networks for mouse and human ESCs we adapted our approach for cell-type-specific data integration and machine learning [1] as follows: 1. Prepared a master list of protein-coding mouse and human gene orthologs with one-to-one homology. 2. Curated a set of training examples focused on ESC self-renewal in pairwise format. Positive/known examples were extracted from recent literature and curated pathways; negative/background pairwise examples were randomly generated from the master gene list, excluding genes involved in positive edges. 3. Collected and standardized high-throughput data from 83 human and 62 mouse ESC studies using diverse experimental techniques (Table 4.1). Used distance/correlation metrics to distill data into pairwise format. 4. Iteratively tested and validated species-specific predictive networks for comparative analysis using the same training set, but with species-specific data compendiums as input.
**Results and Discussion**

We used a cell-type-specific naïve Bayes network methodology (Figure 4.1) [1] to create separate predictive biological networks for hESCs and mESCs focused on self-renewal and closely related biological processes (e.g. pluripotency and cell fate determination). As input data, our Bayes nets take 1) a training set of prior knowledge (also known as a gold standard) comprised of protein-coding gene pairs known to be functionally related (positive training examples) and pairs believed to be unrelated (negative/background training examples), and 2) independent, whole-genome high-throughput datasets (observed evidential data). Based on these inputs, Bayes nets “learn” significant patterns in the evidence, assess data reliability, and then probabilistically predict novel relationships among protein-coding genes based on the most reliable data [56, 58]. (Additional details in Appendix D.) The resulting probabilistic self-renewal networks consisted of 150,363,811 gene pairs that predicted the strength of functional associations among 17,342 protein-coding gene orthologs based on patterns observed in integrated genomic data. Computational evaluations showed that these networks were highly accurate (Supplemental Figure B.6). We used GO term enrichment and network connectivity metrics to assess biological content and functional relevance of our these networks, validate that gene pairs known to significantly influence mESC self-renewal were strongly connected in these probabilistic networks, and identify novel genes with strong functional linkages supported by evidential data.

**hESC Network Predicts Novel Self-Renewal Genes**

Functional associations to known self-renewal genes (SRC scores, see Methods) were calculated for all 17,342 genes in the hESC network and sorted to determine SRC ranks. As expected, many well-studied genes in our training examples are highly ranked by this metric,
Table 4.2. Top Novel hESC Genes Most Strongly Correlated with Self-Renewal

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>Rank</th>
<th>SRC (Scaled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUBB</td>
<td>Tubulin, beta class I</td>
<td>1</td>
<td>0.8264</td>
</tr>
<tr>
<td>HSP90AB1</td>
<td>heat shock protein 90kDa alpha (cytosolic), class B member 1</td>
<td>2</td>
<td>0.8053</td>
</tr>
<tr>
<td>PELI1</td>
<td>pellino homolog 1 (Drosophila)</td>
<td>3</td>
<td>0.7805</td>
</tr>
<tr>
<td>SEMA6A</td>
<td>sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A</td>
<td>4</td>
<td>0.7795</td>
</tr>
<tr>
<td>SLC7A5</td>
<td>solute carrier family 7 (cationic amino acid transporter, y+ system), member 5</td>
<td>5</td>
<td>0.7693</td>
</tr>
<tr>
<td>RGMB</td>
<td>RGM domain family, member B</td>
<td>6</td>
<td>0.7684</td>
</tr>
<tr>
<td>PCBP1</td>
<td>poly(rC) binding protein 1</td>
<td>7</td>
<td>0.7669</td>
</tr>
<tr>
<td>TUBB6</td>
<td>Tubulin, beta 6 class V</td>
<td>8</td>
<td>0.7184</td>
</tr>
<tr>
<td>HSPA4</td>
<td>heat shock 70kDa protein 4</td>
<td>9</td>
<td>0.7653</td>
</tr>
<tr>
<td>SFRP1</td>
<td>secreted frizzled-related protein 1</td>
<td>10</td>
<td>0.7587</td>
</tr>
<tr>
<td>LRRN1</td>
<td>leucine rich repeat neuronal 1</td>
<td>11</td>
<td>0.7574</td>
</tr>
<tr>
<td>PRDM14</td>
<td>PR domain containing 14</td>
<td>12</td>
<td>0.7458</td>
</tr>
</tbody>
</table>

Notes: These genes were identified by selecting the top novel genes (those not involved in positive training set edges) rank ordered by scaled SRC score.

including *POU5F1*, *SOX2*, and *NODAL* (Supplemental Table P.21). The top 12 novel genes (not included in the positive training set) identified by SRC ranks were *TUBB*, *HSP90AB1*, *PELI1*, *SEMA6A*, *SLC7A5*, *RGMB*, *PCBP1*, *TUBB6*, *HSPA4*, *SFRP1*, *LRRN1*, and *PRDM14* (Table 4.2). Literature validation of these genes confirmed that *PRDM14* has recently been shown to play a role in hESC self-renewal and pluripotency [138, 139], and several have been implicated in cancers, including *HSP90AB1*, *SFRP1*, *SEMA6A*, and *PRDM14* [140-143]. Given that β-tubulins serve to define cell shape and are involved in cellular activities that involve motion [144] and have been shown to be involved in positioning the nucleus for cell division and differentiation [145], *TUBB* and *TUBB6* could play a role in hESC colony morphology or regulation of self-renewal through nuclear positioning. Several other highly ranked genes are known to be
involved in embryonic developmental signaling pathways: SFRP1 has been shown to play a key regulatory role in WNT receptor signaling and has been identified as a downstream target of Sonic Hedgehog signaling [146, 147]; SEMA6 has been associated with non-canonical WNT Receptor Signaling and Planar Cell Polarity signaling [142]; and RGMB has been shown to play a regulatory role in BMP signaling [148].

**Human and Mouse ESC Networks are Largely Similar and Significantly Correlated**

To further evaluate our hESC network, we performed a comparative analysis of our new hESC network with an updated version of our previous mESC network [1], generated using the same master list of protein-coding gene orthologs and training set. Differential network analysis shows that the mESC and hESC networks are highly correlated, and most genes share the same functional partners/interactors, as indicated by the posterior probability of edges (Spearman’s rank correlation coefficient $\rho = 0.43$, p-value $< 1 \times 10^{-300}$; Figure4.2a). In particular, positive training set edges were significantly correlated between species (Spearman’s rank correlation coefficient $\rho = 0.4435$, p-value $= 1.08^{-93}$).

Our results confirm the conserved roles of genes known to be involved in early developmental transcriptional regulation and stem cell maintenance in both species, including POU5F1/Pou5f1 (human rank: 1|mouse rank: 1), SOX2/Sox2 (2|5), and NANOG (128|56). The top 1% of genes most conserved in both the mouse and human ESC networks significantly overlap (Hypergeometric p-value= $9.17^{-15}$) and were highly enriched for embryonic-stem-cell-related biological processes, including stem cell maintenance (GO:0010074), negative regulation of cell differentiation (GO:0045596), cell fate commitment (GO:0045165), regulation of transcription, DNA-dependent (GO:0042127), transcription factor activity (GO:0003712) and regulation of gene expression (GO:0010468) using Mus musculus (laboratory mouse) GO
annotations (Table 4.3, Supplemental Tables P.20 and P.21). We observed some species-specific annotation biases [137] in that the same set of genes evaluated using Homo sapiens GO annotations indicated enrichment for transcription factor activity, regulation of transcription, and WNT receptor signaling pathway (GO:0016055).

Table 4.3. Functional Similarities and Differences Between Networks

<table>
<thead>
<tr>
<th>Gene (h/m)</th>
<th>Rank (h/m)</th>
<th>hESC Gene</th>
<th>Rank (h/m)</th>
<th>mESC Gene</th>
<th>Rank (h/m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>POU5F1</td>
<td>Pou5f1</td>
<td>1</td>
<td>1</td>
<td>PEL1</td>
<td>5</td>
</tr>
<tr>
<td>SOX2</td>
<td>Sox2</td>
<td>2</td>
<td>5</td>
<td>LRRN1</td>
<td>15</td>
</tr>
<tr>
<td>NODAL</td>
<td>Nodal</td>
<td>12</td>
<td>114</td>
<td>TUBG1</td>
<td>32</td>
</tr>
<tr>
<td>RIF1</td>
<td>Rif1</td>
<td>14</td>
<td>33</td>
<td>RPL10A</td>
<td>37</td>
</tr>
<tr>
<td>PARP1</td>
<td>Parp1</td>
<td>21</td>
<td>97</td>
<td>RPLP2</td>
<td>38</td>
</tr>
<tr>
<td>LEFTY2</td>
<td>Lefty2</td>
<td>33</td>
<td>85</td>
<td>TALDO1</td>
<td>63</td>
</tr>
<tr>
<td>MYC</td>
<td>Myc</td>
<td>40</td>
<td>38</td>
<td>RPS3A</td>
<td>65</td>
</tr>
<tr>
<td>JARID2</td>
<td>Jarid2</td>
<td>46</td>
<td>11</td>
<td>RPL6</td>
<td>101</td>
</tr>
<tr>
<td>ZFP42</td>
<td>Zfp42</td>
<td>78</td>
<td>27</td>
<td>EIF3A</td>
<td>115</td>
</tr>
<tr>
<td>IFITM1</td>
<td>Ifitm1</td>
<td>107</td>
<td>15</td>
<td>USO1</td>
<td>151</td>
</tr>
<tr>
<td>NANOG</td>
<td>Nanog</td>
<td>128</td>
<td>56</td>
<td>GPM6B</td>
<td>419</td>
</tr>
</tbody>
</table>

Notes: Shared genes strongly associated with ESC self-renewal were identified by finding the intersection of highest-ranking 1% of genes (173 of 17,342) in each network ordered by degree and by SRC in each network. Distinct genes were identified by taking the difference of degree and SRC scores for all genes and selecting the top genes most strongly supported in each network, ranked by SRC.
Figure 4.2. Conservation of Self-Renewal Functional Associations Across Species.
A. Differential network analysis was performed by subtracting posterior probabilities of pairwise functional relationships in the hESC network from those in the mESC network. Although the majority of edges were highly conserved, there are notable differences between species localized to the extreme tails of the distribution (see Supplemental Table S10 and Supplemental File S5). B. “Core” self-renewal genes identified from our positive training examples (i.e. those most tightly associated with each other) significantly overlap between species (Hypergeometric p-value = 8.56e-7).
As described in Materials and Methods, we identified 173 core self-renewal genes in the mESC network and 111 core self-renewal genes in the hESC network with a significant overlap of 73 shared genes (Figure 4.2B; Supplemental Table P.21; Hypergeometric p-value = 8.56^{-7}). As expected, well-studied genes known to play a significant role in self-renewal and pluripotency across species were ranked highest in both networks. These key players included transcriptional regulators POU5F1/Pou5f1 (hESC rank: 1 | mESC rank: 1), SOX2/Sox2 (2 | 5), and NANOG/Nanog (128 | 56) (Table 4.3, Supplemental Table P.22). Highly conserved novel genes (not included in training set examples) included PARP1/Parp1 (21 | 97), a regulator of chromatin structure, transcription, and DNA damage repair that has recently been shown to be required for reprogramming mouse embryonic fibroblasts (MEFs) to a pluripotent state [149, 150]; and IFITM1/Ifitm1 (107 | 15), an interferon-induced transmembrane program identified as a downstream target of WNT receptor signaling during gastrulation involved in somitogenesis and mesoderm formation [151].

In our training examples, we included all components of KEGG signaling pathways known to influence early development and ESC cell fate, such as FGF and JAK/STAT signaling [129, 152, 153]. However, as these pathways include many homologous components, it is likely that only a subset of documented pathway participants are active in ESCs, and some interactions may be species specific (e.g. LIF-Activated JAK/STAT signaling in mESCs, and ACTIVIN-A-activated FGF signaling in hESCs). As such, the bulk of training set genes with low SRCs in both networks consist of signaling pathway participants that appear to be more likely involved in signaling in a cellular context other than ESCs, such as FGF22/Fgf22 (15800 | 14199), WNT10A/Wnt10a (12140 | 10554), and INHBA/Inhba (12583 | 14301). Surprisingly, given its role in X-chromosome inactivation [130, 154], XIST/Xist (13765 | 11727) ranked poorly in both networks. This could be because, while it is measured by most human and mouse microarray
platforms, it may not correlate with other specific gene expression regulatory and protein-DNA binding genes that are strongly associated with self-renewal. While our compendiums contain extensive high-throughput data from a broad range of experimental techniques (Table 1), the bulk of our data for both species is derived primarily from gene expression studies (using microarray or RNA-Seq technologies) and from protein-DNA binding data (from ChIP-Chip and ChIP-Seq assays). While we show that these data can be highly informative for elucidating protein function in a specific cellular context, the scarcity of additional high-throughput data types, such as protein-protein-binding assays performed in ESCs limits our approach to identifying genes related primarily through expression or direct transcriptional regulation.

**Predictive Networks Emphasize Important Signaling Pathway Differences**

Although the mESC and hESC networks are highly correlated, suggesting that gene functional associations in the context of ESC self-renewal are largely conserved, there are striking differences (Table 3, Supplemental Table P.23). Divergent functional linkages are particularly evident in developmental signaling pathways and our predictive networks recapitulate many known differences in signaling cues that prompt naïve mouse versus primed human ESCs to sustain or exit a self-renewing state. To visualize these differences, we created predictive mESC and hESC developmental signaling pathways using gene elements identified in sub-sections of KEGG pathways, overlaid with our predicted connection probabilities and functional correlations to self-renewal (SRCs) for those genes (Figures 4.3 and 4.4).
Figure 4.3. Predicted FGF Signaling Pathway Relationships Across Species. These network models show the predicted strength of relationships between all possible FGF signaling ligand-receptor pairs. Gene nodes are colored by SRC score (light gray = weak correlation to self-renewal; dark blue = strong self-renewal correlation) and edge color/thickness indicates the strength of predicted functional association between ligand and receptor (yellow/thin = weak; teal/thick = strong). 

A. The hESC FGF signaling model shows that the most strongly connected ligand-receptor pairs with high-ranking SRCs are FGF2 and FGF19 (the human ortholog of Fgf15), which are most strongly associated with FGFR1 and FGFR3. FGF2 is known to activate FGF signaling in hESCs.

B. The mESC FGF Signaling model showed that the most strongly connected ligand-receptor pairs with high ranking SRCs are Fgf4 and Fgf5, which are both associated with Fgfr1 and Fgfr2, and Fgf15 which is most strongly connected to Fgfr2. In mESCs, Fgf4 is known to activate FGF signaling, Fgf5 is associated with FGF signaling in the late stage blastocyst and epiblast, and Fgf15 has been shown to be involved in early neurodevelopment.
**Figure 4.4.** Predicted JAK/STAT Signaling Pathway Relationships Across Species. Analogous to Figure 3, species-specific predictions are shown for a portion of the KEGG pathway map for JAK/STAT signaling. **A.** The hESC predicted JAK/STAT signaling pathway showed that Stat3 is the gene most correlated to self-renewal, while upstream pathway participants exhibited lower SRC scores, suggesting that cytokines other than the IL-6 family, or signaling cross talk, may be required for STAT3 activation in hESCs. **B.** The mESC predicted JAK/STAT signaling pathway showed a strong connection between Lif and its putative target, Il6st. Further, key known mouse self-renewal genes in this pathway, including Lif, Il6st, and Stat3 are strongly correlated to other self-renewal genes as indicated by SRC score.
For example, FGF signaling has been shown to play an important role in regulating ESC self-renewal and differentiation (as well as myriad other processes) and may be involved in promoting the ESC transition from a naïve to primed pluripotent state. Mice and humans have 22 FGF ligands and 5 FGFR receptors, each with specific expression patterns that change over time during early development, and different ligand-receptor pairs have varying functional roles depending on the cellular context. In mESCs, FGF4-activated ERK signaling promotes differentiation [155], whereas in hESCs, FGF2-activated ERK signaling sustains self-renewal [29, 155]. We compared functional associations among FGF ligand-receptor pairs using our hESC and mESC networks and found our predictions clearly recapitulate known differences between species (Figure 4.3). The FGF2/Fgf2 (48|14300) ligand was strongly correlated to self-renewal in the hESC network only; while FGF4/Fgf4 (3797|156) and FGF5/Fgf5 (2965|308) ranked highest in the mESC network only. In both species, these active FGF ligands were strongly linked to the FGFR1/Fgfr1 (126|302) and FGFR2/Fgfr2 (198|379) receptors, while the FGFR3/Fgfr3 (170|6063) and FGFR4/Fgfr4 (1191|7560) receptors were strongly linked only in the human network. In addition, both our hESC and mESC networks showed high-ranking SRCs for the FGF19/Fgf15 (144|622) ligand, though in humans it was associated with the FGFR1/Fgfr1 (126|302) and FGFR3/Fgfr3 (170|6063) receptors, while in mice this ligand was associated primarily with FGFR2/Fgfr2 (198|379). In mESCs, Fgf15 mediates differentiation to the neuroectoderm, while the function of the human ortholog FGF19 remains less clear, although it has been implicated in general neuronal development [156]. Through the similar functional linkage patterns of FGF19/Fgf15 in the hESC and mESC networks, our models predict FGF19 may play a similar role in human neural development as Fgf15 does in the laboratory mouse (Figure 4.3B).
We also examined LIF-activated JAK/STAT signaling, as LIF promotes self-renewal by activating the JAK/STAT3 and PI3K/AKT signaling pathways in mESCs, but is not required by hESCs to sustain self-renewal [129, 152]. While STAT3/Stat3 (202|485) is strongly associated with self-renewal in both species in our networks, it appears that the mode of regulation may be diverged. This difference is clearly supported in our probabilistic pathways focused on JAK/STAT signaling stimulated by the Interleukin-6 (IL-6) family of cytokines (Figure 4.4) [157]. In the mESC network, a strong connection exists between the LIF/Lif (3457|1317) cytokine and its well-characterized signal transducer IL6ST/Il6st (also known as gp130; 9107|347), which in turn is tightly linked to tyrosine kinase JAK3/Jak3 (10600|3126) and then on to STAT3/Stat3 (202|485) (Figure 4.4A). These predicted linkages reflect the most well-documented LIF-activated IL-6 JAK/STAT signaling cascade in mESCs [158-160]; however these strong associations are not present in our hESC network. Our mESC results also showed functional linkages between LIF/Lif and the IFNGR2/Ifngr2 (5801|946) receptor, which associates most strongly with JAK1/Jak1 (2789|3808) and JAK2/Jak2 (6625|1989), both of which strongly link to STAT3/Stat3. In contrast, within the confines of the KEGG JAK/STAT pathway, the hESC network predicts that LIF/Lif and the many other IL-6 cytokines are not strongly associated with self-renewal (based on SRC scores), and shows only weak connections between LIF and the Il6ST signal transducer [19] (Figure 4.4B). Rather, in hESCs, the cytokines IL11/I11 (2939|15541) and LIF were moderately associated with the IFNGR1/Ifngr1 (2701|6655) and OSMR/Osmr (2060|8138) receptors, which were moderately linked to the JAK1/Jak1 (2789|3801) and TYK2/Tyk2 (11394|8747) tyrosine kinases, respectively (note that while Tyk2 is annotated in the mouse JAK/STAT pathway by KEGG, it is not listed as a participant in the human KEGG pathway). Global expression profiling in mESCs has shown that the interferon gamma receptor IFNGR1/Ifngr1 is regulated by POU5F1/Pou5f1 [161], and although not documented as active in
early embryonic development, *OSMR|Osmr* is predicted to have a strong functional association with *POU5F1* supported by both gene expression and protein-DNA-interaction data in the human network. In contrast to these moderate associations with *STAT3*, our hESC network predicts strong *STAT3* associations with members of other signaling pathways, including *TCF7L1*, *FGF2*, *FGFR1*, and *BMPR2*, suggesting potentially significant cross-talk and/or alternate modes of *STAT3* regulation. However, these strong associations are not observed for *Stat3* in our mESC network, where *Stat3* is most strongly associated with known LIF and JAK/STAT signaling pathway members, such as *Fos*, *Pim1*, and *Spry4*.

**Metabolic Differences Between Species Highlighted in Predictive Networks**

One of the most striking differences between our mouse and human ESC networks concerns threonine catabolism, which is required for mESC self-renewal, likely through the *TDH|Tdh* (L-threonine dehydrogenase) gene, which supports accelerated cell cycle kinetics by catabolizing threonine into glycine and acetyl-CoA, which is used by the TCA cycle to generate ATP [87, 119]. In our mouse network, *TDH|Tdh* (11393|26) has a strong correlation to mESC self-renewal and is tightly linked to many core self-renewal genes, including *POU5F1|Pou5f1* (1|1), *SOX2|Sox2* (2|5), *RIF1|Rif1* (14|33), and *NR0B1|Nr0b1* (5583|487). The functional relationship between *Tdh* and these genes is largely supported by ChIP-chip binding data from studies investigating the regulatory circuitry of mESCs and microarray data from a study analyzing mESC differentiation (Figure 4.5B). In contrast, *TDH* is not correlated to self-renewal in our hESC network and not tightly connected to any genes in our positive training set (Figure 4.5A). Differential and functional analyses of the mESC and hESC networks did not reveal a direct metabolic equivalent to threonine dehydrogenase in humans. Literature validation showed that the human *TDH* gene that encodes threonine dehydrogenase has been rendered
non-functional due to 3 mutations (2 AG-to-GG splice acceptor mutations in Exons 4 and 6, and a nonsense mutation in Exon 6) [87]. However, hESCs grow at a slower rate than mESCs (with a doubling time of 35 hours as opposed to every 4-5 hours), and it is not yet known if the difference in growth rate might be due to the absence of TDH or if there may be some selective advantage for inactivating TDH in humans [87, 119, 162].

Figure 4.5. Differences in mESC and hESC Threonine Metabolism. We used our StemSight Scout data visualization tool to create comparative network views centered around L-threonine dehydrogenase (TDH), which supports accelerated cell cycle kinetics in mESCs, but is not functional in hESCs. In this view, Scout displays a maximum of 30 edges per genes with a minimum edge weight of 0.5 between gene pairs. Node and edge colors are as in Figures 3 and 4, except that edges contained in the positive training set are colored orange. A. The hESC TDH-centric network shows that TDH is weakly correlated to genes in our training set and has no strong functional associations to known self-renewal genes. B. The mESC Tdh-centric network illustrates that Tdh is strongly correlated to self-renewal genes and has strong predicted functional associations with known self-renewal genes, including Pou5f1, Sox2, Nr0b1, Klf2, Zfp42, Gdf3, and Fbx015.
Comparative Network Analyses Reveal Novel Species-Specific Differences

To discover novel species-specific differences, we focused on the top 0.001% (1503) of gene pairs with the greatest difference between the mESC versus hESC networks (Supplemental Table P.23). There were 86 genes involved in edges strongly supported only in the mESC network, including genes annotated to chordate embryonic development (GO: 0043009), stem cell maintenance (GO: 0019827), negative regulation of cell differentiation (GO: 0045596), and regulation of transcription, DNA-dependent (GO: 0042127). In contrast, 179 genes were involved in edges supported only by the hESC network, including genes annotated to WNT Receptor Signaling (GO: 0016055), MAPK Cascade (GO: 0000165), FGF Signaling (GO: 0008543), ATP-Binding (GO: 0005524), and cell cycle regulation (GO: 0051726), emphasizing the different set of developmental signaling cues observed in early cell fate decisions in hESCs. These key differences were echoed when we compared the top 1% of novel genes ranked by SRC (the 173 highest ranking genes not included in our positive training set examples). Top-ranked novel mESC genes were predominantly associated with differentiation and transcriptional regulation processes, whereas the novel hESC genes were largely related to regulation of translation (GO: 0006417), microtubule-based process (GO: 0007017), regulation of cell cycle (GO: 0051726), cell division (GO: 0051301), and cell adhesion (GO: 0007155). Although there are known differences in cell cycle controls that permit rapid cell cycling in ESCs, it is not yet known whether the cell cycle regulates pluripotency. Cell cycle length is believed to not determine pluripotency, because different pluripotent cell types divide at different rates [163]; however, the enrichment of cell cycle component genes in our novel hESC gene lists suggest that there may be functional differences in our mouse and human networks that can be mined to improve our understanding of how the core cell cycle machinery adapts to the timing requirements of primed and naïve ESCs. Interestingly, many of the genes most strongly correlated to self-renewal in the hESC
network only (Table 4.3) are involved in the translational regulation of protein synthesis, such as the ribosomal proteins RPL10A|Rpl10a (37|8606), RPLP2|Rplp2 (38|7284), RPS3A|Rps3a (65|11303), and RPL6|Rpl6 (101|11695), as well as the translation initiation factor EIF3A|Eif3a (115|12160). Intriguingly, genome-wide studies using hESCs have shown that LIN28|Lin28a (62|224), a known regulator of self-renewal, targets include these ribosomal and translation-supporting genes, which are important for growth and survival [164]. In contrast, Lin28a targets in mESCs are enriched for translational repressors of membrane-bound proteins, secretory proteins, and proteins destined for the endoplasmic reticulum/Golgi lumen [165], suggesting an alternative regulatory role between species for LIN28|Lin28a, despite supporting self-renewal in both species.

**Comparative Network Visualization Promotes Novel Gene Discovery**

To make our probabilistic comparative networks readily available to the stem cell research community, we provide interactive, online visualization resources at www.StemSight.org that can be used to view the hESC and mESC networks independently or comparatively. StemSight Scout, implemented using ThinkMap visualization technology, highlights potentially novel self-renewal genes by coloring nodes based on their SRC score and illustrates the weight of predicted interactions by coloring edges based on the inferred posterior probabilities. Edges included in our training set examples are color-coded, making it easy to visually segregate novel from known. For positive training set edges, links are provided to the original articles documenting the relationship, while for novel edges, the evidential data supporting that functional relationship is displayed with links to the study in PubMed.
The comparative view enables researchers to search for interactions among one or more genes in both networks, and displays results in a convenient side-by-side window for direct comparisons (Figure 5), and is especially helpful for contrasting connectivity of gene hubs and pathway participants in the hESC versus mESC networks.

**Conclusions**

Determining the molecular underpinnings of stem cell self-renewal and cell fate determination is vital for understanding and refining the production of iPS cells, and has potential implications for targeting cancer stem cells and demystifying aging processes in adult stem cells. Our results confirm that multiple signaling pathways contribute to the balance between self-renewal and differentiation in ESCs, and further suggest that these complex pathways should not be considered monolithically, but recognized as general mechanisms whose details and interactions are only partially understood. For example, while our results indicate that STAT3/Stat3 is strongly associated with self-renewal in both humans and mice, upstream regulation of signaling appears significantly diverged between species. Similar observations and conclusions can be drawn from our results for various members of WNT, NODAL/ACTIVIN-A-activated FGF, TGF-β, and many other signaling and metabolic pathways.

Our computational integration of hundreds of high-throughput datasets provides an entry point to the greater understanding of these pathways. For example, our evaluation of STAT3 interactions suggests potentially significant crosstalk and alternate modes of human STAT3 regulation. As STAT3/Stat3 is a multi-functional gene implicated in phenotypes and disorders ranging from circadian rhythms to cancer, developing cancer therapies targeting the STAT3 pathway requires investigation of potential off-target effects, which in turn requires
expanding our understanding of its context-specific regulation. Thus, unraveling the diversity of systems-level interactions and crosstalk is vital for the use of model organisms, and is likely key for dissecting the cell type-specific roles of multi-functional genes within humans. Similarly, the divergent role of \(TDH/Tdh\) between species in our networks likely reflects the distinct metabolic mechanisms of these cell types. While \(Tdh\) supports increased growth rates in mESCs, it is possible that the increased activity of translational regulatory genes in hESCs (such as the ribosomal and translation initiation genes identified only in our human networks) play a similar role in a human context.

By focusing here on human and mouse ESC self-renewal, we have developed methods for cross-species and cross-context comparisons to assist stem cell researchers in evaluating known and novel genes involved in ESC self-renewal and differentiation. In the future, comparative networks for additional stem cell types, including mEpiSCs, iPSCs, ESCs from other species, as well as adult stem cells. Primed mEpiSCs would be particularly useful for investigating differences between the naïve and primed pluripotent state within mouse, as well as for comparing primed pluripotent cells in different species.

While computational systems biology efforts such as ours begin to scrutinize complex molecular interactions, additional laboratory efforts are required to confirm these hypotheses and to determine precise mechanisms. The data generated by these efforts can be re-incorporated into machine learning efforts to refine our training sets, provide additional input evidential data, and ultimately improve our understanding of context-specific, whole-genome interactomes. As such, the full results of this study are freely available to the stem cell community at www.stemsight.org, where users can explore our hESC and mESC self-renewal networks to place their results into a broader context, draw new hypotheses, or prioritize candidate genes.
Acknowledgments

This project was partially funded by The Jackson Laboratory. KGD is supported by a Pharmaceutical Researchers and Manufacturers of America (PhRMA) Foundation Fellowship in Informatics, and has received funding from the University of Maine Graduate School of Biomedical Sciences, and the National Science Foundation Integrated Graduate Education and Research Traineeship (IGERT) program grant 0221625 to conduct this work. MAH is an Ellison Medical Foundation New Scholar in Aging; he is partially supported by National Institutes of Health /National Institute of Arthritis and Musculoskeletal and Skin (NIH/NIAMS) grant R21 AR069981-01 (Hibbs, Ackert-Bicknell), and National Institutes of Health (NIH) Grant P50 GM076468-06 (Churchill).
CHAPTER 5

CONCLUSIONS AND FUTURE WORK

This study shows that cell-type-specific predictive networks created using Bayesian network data integration and machine learning techniques are a powerful computational resource for exploring mouse and human pluripotent stem cell biology in the context of self-renewal and early cell fate determination. My pluripotent stem cell networks, generated using a manually curated gold standard focused on self-renewal and closely related biological processes coupled with consistently processed evidential data sets limited to a single cell type in one species, are computationally robust and biologically relevant. Through my mESC and hESC predictive networks, I identified many novel genes functionally linked to self-renewal or related biological processes that are currently being experimentally validated by collaborators. My comparative mESC and hESC predictive networks reveal interesting contrasts that highlight species-specific differences in functional gene associations that may influence the pluripotent cell state, energy metabolism, and morphology in mouse and human ESCs. These networks may be mined for meaningful and testable insights into novel ESC biology and used to identify conserved and divergent functional associations among genes across species.

Although this work focused on mouse and human embryo-derived pluripotent stem cells, there are many additional cell types and tissues that could be explored using this type of approach. One logical next step for this study would be to create predictive networks for other types of pluripotent cells, such as mEpiSCs, and pluripotent stem-like cells, such as mouse and human iPSCs. With mEpiSCs, the biggest challenge is the lack of available high-throughput data for these cells. While there was a surge of interest in mEpiSCs soon after they were first isolated and shown to have characteristics similar to hESCs [24], their perceived research value was soon
eclipsed in the field by the discovery of iPSCs. Although I did create pilot predictive networks for mEpiSCs using a mESC self-renewal gold standard, there was insufficient high-throughput data available to create meaningful biological networks. What little data I could collect was largely from gene expression studies, and therefore provided information only on one facet of functional associations within the cell. Creating predictive networks for iPS cells presents a different challenge in that there are many methods for reprogramming different types of adult cells to a pluripotent-like state. Therefore selection of appropriate evidential data could be difficult and time-consuming. Anecdotally, researchers have claimed that with sufficient cell culture passages, iPS cells can be fully reprogrammed to a naïve pluripotent state, but this raises questions about adaptation in the dish [166], and not all studies consistently report the passage number for each sample measured. Even so, consensus predictive networks could provide useful insights into differences between ESCs and iPS cells within a single species. Alternatively, consensus ESC networks could potentially be used as background for mining more targeted datasets using iPS cells from a single laboratory.

Other potential biological applications for this methodology would be to create context-specific predictive networks for lineage-specific stem cells and related cancer stem-like cells, such as HSCs and leukemias, or NSCs and brain cancers. The limiting factors in such studies are simply the amount of evidential data available and the time required to assemble a context-specific gold standard. After ESCs, hematopoietic stem cell lineages and related cancers of the blood are the most well-studied and characterized, and would be a promising subject to explore. Additionally, my approach could be used to create predictive networks of virtual co-cultures that model functional associations among multiple cells that comprise a specific tissue. Theoretically, this type of tailored naïve Bayes classification method could be applied to any well-studied,
focused biological problem for which there is sufficient training examples and high-throughput data.

Challenges of Study and Platform Bias in Comparative Studies

One of the key lessons learned in this work is that appropriate data selection is critical for creating biologically meaningful and testable predictive networks. One of the biggest hurdles of creating comparative networks using different cell types or the same cell types from different species is that data and study bias varies by field and research subspecialties. For example, when assembling data for our comparative mESC and hESC data, I observed that most studies conducted using mESCs included conditions that perturbed expression or binding activity of canonical transcription factors: Pou5f1, Sox2, and Nanog. This study bias was not as noticeable in hESC data. Similarly, at the time of writing, I was able to find more epigenetic profiling data and ENCODE data using human ESC cell lines than for mouse. Although the resulting networks created using separate data compendiums composed of datasets biased in different ways were still biologically useful, these types of biases do affect predictions and edge weights, making it harder to do one-to-one comparisons.

Opportunities for Enhancements: StemSight Release 2.0

This study serves to extend the life of high-throughput data by integrating it into consensus predictive networks that are freely available to the stem cell community. Unlike other computational studies focused on ESCs, this work not only provides access to network files containing millions of edges, but also bundles together visualization and analysis tools specifically designed for mining these networks. My predictive network visualization tool,
StemSight Scout, provides dynamic, responsive options for visually exploring large complex networks online without requiring users to download and install additional software. It provides many features not available in other similar visualization tools, and the StemSight development team has been working hard on a new release of Scout for HTML5 that will further enhance network visualizations and eliminate pesky security and display issues associated with JAVA-based applets. New functionality planned for Scout Release 2.0, expected to be available in September 2013, will include multi-centered gene search capabilities that will allow users to fetch network views for multiple genes (e.g. Tdh, Pou5f1, Sox2, and Rif1) in one query, as well as a comparative network interface that enables users to view query results for two networks (mESC and hESC) in the same browser window.

Additional refinements to my methodology for creating predictive networks could include alternative methods for developing lists of gene orthologs that permit one-to-many associations among protein-coding gene orthologs, new methods for mapping edge associations to gene function that make inference scores (edge weights) and correlations to biological functions (e.g. SRCs), easier for biologists to interpret. Ultimately, I would like to incorporate directed edges into our weighted networks as a first step toward predicting mechanisms, and provide predictions for associations among genes products and other cellular elements, such as RNAs and metabolites. I expect that ongoing feedback from the stem cell and data visualization communities will provide additional suggestions for further enhancing the tools and methods described in this work.
# REFERENCES


76. Markowetz F, Spang R: **Inferring cellular networks--a review.** *BMC Bioinformatics* 2007, **8** Suppl 6:S5.


88. Ryu JM, Han HJ: **L-threonine regulates G1/S phase transition of mouse embryonic stem cells via PI3K/Akt, MAPKs, and mTORC pathways.** *J Biol Chem* 2011, **286**(27):23667-23678.


APPENDIX A

GLOSSARY OF TERMS AND ABBREVIATIONS

The fields of stem cell biology and machine learning use many descriptive and technical terms, acronyms and abbreviations. This Appendix defines frequently used stem cell biology and machine learning terms, and provides full names for developmental signaling and metabolic pathways, experimental techniques and other terms abbreviated in this work. It also provides a partial listing of gene symbols referenced that including MGI systematic IDs, official gene symbol, and full gene name. Complete lists of genes referenced in this study are available in Appendix E, Supplemental Tables P.12 and P.14.

Stem Cell and Developmental Biology Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem cell</td>
<td>Cells that have the capacity to divide repeatedly to produce more cells of the same type (self-renewal) as well as one or more types of specialized cells (differentiation) that perform more specific functions in developing or adult organisms.</td>
</tr>
<tr>
<td>Self-renewal</td>
<td>The ability of a stem cell to divide and produce (give rise to) daughter cells</td>
</tr>
<tr>
<td>Differentiation</td>
<td>The ability of a stem cell to produce or differentiate into one or more cells types committed to perform a more specialized function.</td>
</tr>
<tr>
<td>Proliferation</td>
<td>An increase in the number of cells of the same type through cell division.</td>
</tr>
<tr>
<td>Potency or Potential</td>
<td>Defines the limits of a stem cell’s ability or potential to generate one or more specialized cell types.</td>
</tr>
<tr>
<td>Give rise</td>
<td>A common term for stem cell differentiation used when defining which types of differentiated cells a stem cell has the potential to produce.</td>
</tr>
<tr>
<td>Cleavage</td>
<td>A series of rapid mitotic divisions that occur during early embryonic development, immediately following fertilization of the egg. During cleavage, the volume of the egg cytoplasm is divided into an increasing number of smaller, nucleated cells called blastomeres. In mammalian embryogenesis, the fertilized egg divides or cleaves into two cells, then four cells, then eight cells, all with the same potential. At the end of the cleavage stage, the blastomeres form a sphere of more specialized cells called the blastocyst.</td>
</tr>
</tbody>
</table>
Germ layers

After the rate of mitotic division that occurs during cleavage slows down, blastomeres undergo a process called gastrulation, during which the individual cells rearrange themselves into a structure called a gastrula and take on more specialized functions. There are three distinct layers of cells in the gastrula, each of which contribute to different functional parts of the developing organism: the endoderm, ectoderm, and mesoderm. Pluripotent stem cells have the potential to differentiate into any cells that form these three germ layers and are therefore said to give rise to all adult cell types.

Totipotent

A stem cell with the capacity to differentiate into all cell types in a developing organism, including the embryo and fetal placenta (trophectoderm). Totipotent stem cells have limitless replicative abilities, but occur only transiently in nature, during the post fertilization cleavage stage of embryogenesis. In Latin, totipotent means “capable of forming everything.”

Pluripotent

A stem cell with the capacity to differentiate into all three germ layers of the developing embryo. Pluripotent stem cells occur only transiently in nature, during the late cleavage (pre- and post-implantation blastocyst) phases of embryogenesis, prior to gastrulation. In Latin, pluripotent means “capable of forming many things.”

Multipotent

A stem cell with the capacity to differentiate into a limited number of specialized cell types. For example, hematopoietic stem cells can produce multiple types of blood cells.

Unipotent

A stem cell with the capacity to differentiate into one specialized cell type. Skin epithelial cells are an example of a highly proliferative unipotent stem cell.

Inner Cell Mass (ICM)

A cluster of pluripotent cells found inside the blastocyst during the cleavage stage of early embryogenesis. Cells in the inner cell mass give rise to the developing embryo.

Trophectoderm

Cells that form the surface of the blastocyst (also called trophoblasts). These cells do not produce embryonic structures, but form the embryonic portion of the placenta. The formation of the trophectoderm is the first differentiation event in mammalian development.

Embryonic stem cell (ESC)

Pluripotent stem cells that comprise the inner cell mass of the blastocyst, with the potential to give rise to all three germ layers in the developing embryo, thereby contributing to all cell types.

Induced Pluripotent Stem Cell (iPSC)

Adult cells that have been genetically reprogrammed to a pluripotent-stem-cell-like state. Like embryo-derived pluripotent cells, iPSCs express stem cell markers and form tumors containing cells from all three germ layers. Although there are now many methods for inducing this pluripotent-like state in different types of adult cells, the first iPSCs were created using fibroblasts transfected with reprogramming Yamanaka factors (Pou5f1, Sox2, Myc, Klf4) via viral vectors.
**Machine Learning**

**Computer Agent**  Something that operates autonomously, perceives its environment, persists over prolonged periods of time, adapts to change, and pursues goals

**Machine Learning**  When an agent improves its performance on future tasks after making observations. Machine learning can be applied to any problem where a pattern exists that can’t be pinned down mathematically and data exists to learn from. There are many analytical approaches to machine learning and a wide range of practical applications, from spam filters to optical character recognition.

**Supervised Learning**  When an agent observes examples input-output pairs and learns a function that maps from input to output.

**Bayesian Learning**  When an agent calculates the probability of each of each hypothesis, given the data, and makes predictions using all hypotheses, weighted by their probabilities, as opposed to using a single “best” hypothesis.

**Bayesian Network**  A graphical structure that represents dependencies among variables for any joint probability distribution.

**Naïve Bayes Network**  A probability distribution where a single cause (prior) directly influences a number of effects (observations), all of which are conditionally independent

**Developmental Signaling and Metabolic Pathways**

<table>
<thead>
<tr>
<th>Pathway Reference</th>
<th>Full Pathway Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAK/STAT Signaling</td>
<td>Janus Kinase/Signal Transducer and Activator of Transcription Signaling Pathway</td>
</tr>
<tr>
<td>WNT Signaling</td>
<td>Wingless-related MMV Integration Site (WNT) Growth Factor Family Signaling Pathway</td>
</tr>
<tr>
<td>FGFR Signaling</td>
<td>Fibroblast Growth Factor Receptor Signaling Pathway</td>
</tr>
<tr>
<td>MAPK Signaling</td>
<td>Mitogen Activated Protein Kinase Signaling Pathway</td>
</tr>
<tr>
<td>TGF-β Signaling</td>
<td>Transforming Growth Factor Beta Signaling Pathway</td>
</tr>
<tr>
<td>TCA cycle</td>
<td>tricarboxylic acid cycle (also known as the citric acid cycle)</td>
</tr>
</tbody>
</table>
Experimental Techniques and Biological Elements

- acetyl-CoA: Acetyl coenzyme A
- AP-MS: Affinity Purification followed by Mass Spectrometry
- ChIP-Chip: Chromatin Immunoprecipitation followed by microarray technology (chip); also known as ChIP-on-chip
- ChIP-Seq: Chromatin Immunoprecipitation sequencing
- RNAi: RNA interference
- shRNA: Small hairpin RNA or Short hairpin RNA

Gene Symbol Reference

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>MGI ID</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abcc4</td>
<td>MGI:2443111</td>
<td>ATP-binding cassette, sub-family C (CFTR/MRP), member 4</td>
</tr>
<tr>
<td>Akap12</td>
<td>MGI:1932576</td>
<td>A kinase (PRKA) anchor protein (gravin) 12</td>
</tr>
<tr>
<td>Anp32a</td>
<td>MGI:108447</td>
<td>acidic (leucine-rich) nuclear phosphoprotein 32 family, member A</td>
</tr>
<tr>
<td>Brca1</td>
<td>MGI:104537</td>
<td>breast cancer 1</td>
</tr>
<tr>
<td>Cbx2</td>
<td>MGI:88289</td>
<td>chromobox 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2</td>
</tr>
<tr>
<td>Cited2</td>
<td>MGI:1306784</td>
<td></td>
</tr>
<tr>
<td>Creb3</td>
<td>MGI:99946</td>
<td>cAMP responsive element binding protein 3</td>
</tr>
<tr>
<td>Crmp1</td>
<td>MGI:107793</td>
<td>collapsin response mediator protein 1</td>
</tr>
<tr>
<td>Dbf4</td>
<td>MGI:1351328</td>
<td>DBF4 homolog (S. cerevisiae)</td>
</tr>
<tr>
<td>Dnmt3l</td>
<td>MGI:1859287</td>
<td>DNA (cytosine-5')-methyltransferase 3-like</td>
</tr>
<tr>
<td>Etv5</td>
<td>MGI:1096867</td>
<td>ets variant gene 5</td>
</tr>
<tr>
<td>Fzd2</td>
<td>MGI:1888513</td>
<td>frizzled homolog 2 (Drosophila)</td>
</tr>
<tr>
<td>Fzd5</td>
<td>MGI:108571</td>
<td>frizzled homolog 5 (Drosophila)</td>
</tr>
<tr>
<td>Gbx2</td>
<td>MGI:95668</td>
<td>gastrulation brain homeobox 2</td>
</tr>
<tr>
<td>Gjb3</td>
<td>MGI:95721</td>
<td>gap junction protein, beta 3</td>
</tr>
<tr>
<td>H2afx</td>
<td>MGI:102688</td>
<td>H2A histone family, member X</td>
</tr>
<tr>
<td>Ifitm1</td>
<td>MGI:1915963</td>
<td>interferon induced transmembrane protein 1</td>
</tr>
<tr>
<td>Il6st</td>
<td>MGI:96560</td>
<td>interleukin 6 signal transducer</td>
</tr>
<tr>
<td>Ina</td>
<td>MGI:96568</td>
<td>internexin neuronal intermediate filament protein, alpha</td>
</tr>
<tr>
<td>Jam2</td>
<td>MGI:1933820</td>
<td>junction adhesion molecule 2</td>
</tr>
<tr>
<td>Jarid2</td>
<td>MGI:104813</td>
<td>jumonji, AT rich interactive domain 2</td>
</tr>
<tr>
<td>Klf4</td>
<td>MGI:1342287</td>
<td>Kruppel-like factor 4 (gut)</td>
</tr>
<tr>
<td>Klf9</td>
<td>MGI:1333856</td>
<td>Kruppel-like factor 9</td>
</tr>
<tr>
<td>Lif</td>
<td>MGI:96787</td>
<td>leukemia inhibitory factor</td>
</tr>
<tr>
<td>Gene</td>
<td>MGI</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Lifr</td>
<td>MGI:96788</td>
<td>leukemia inhibitory factor receptor</td>
</tr>
<tr>
<td>Lpp</td>
<td>MGI:2441849</td>
<td>LIM domain containing preferred translocation partner in lipoma</td>
</tr>
<tr>
<td>McI1</td>
<td>MGI:101769</td>
<td>myeloid cell leukemia sequence 1</td>
</tr>
<tr>
<td>Mkrm1</td>
<td>MGI:1859353</td>
<td>makorin, ring finger protein, 1</td>
</tr>
<tr>
<td>Msh6</td>
<td>MGI:1343961</td>
<td>mutS homolog 6 (E. coli)</td>
</tr>
<tr>
<td>Msx2</td>
<td>MGI:97169</td>
<td>homeobox, msh-like 2</td>
</tr>
<tr>
<td>Myst2</td>
<td>MGI:2182799</td>
<td>MYST histone acetyltransferase 2</td>
</tr>
<tr>
<td>Nanog</td>
<td>MGI:1919200</td>
<td>Nanog homeobox</td>
</tr>
<tr>
<td>Nolc1</td>
<td>MGI:1918019</td>
<td>nucleolar and coiled-body phosphoprotein 1</td>
</tr>
<tr>
<td>Nr0b1</td>
<td>MGI:1352460</td>
<td>nuclear receptor subfamily 0, group B, member 1</td>
</tr>
<tr>
<td>Pdc</td>
<td>MGI:98090</td>
<td>phosducin</td>
</tr>
<tr>
<td>Phc1</td>
<td>MGI:103248</td>
<td>polyhomeotic-like 1 (Drosophila)</td>
</tr>
<tr>
<td>Plcg2</td>
<td>MGI:97616</td>
<td>phospholipase C, gamma 2</td>
</tr>
<tr>
<td>Pou5f1</td>
<td>MGI:101893</td>
<td>POU domain, class 5, transcription factor 1</td>
</tr>
<tr>
<td>Rb1</td>
<td>MGI:97874</td>
<td>retinoblastoma 1</td>
</tr>
<tr>
<td>Reep3</td>
<td>MGI:88930</td>
<td>receptor accessory protein 3</td>
</tr>
<tr>
<td>Rhob</td>
<td>MGI:107949</td>
<td>ras homolog gene family, member B</td>
</tr>
<tr>
<td>Rif1</td>
<td>MGI:1098622</td>
<td>Rap1 interacting factor 1 homolog (yeast)</td>
</tr>
<tr>
<td>Rpp25</td>
<td>MGI:2143151</td>
<td>ribonuclease P 25 subunit (human)</td>
</tr>
<tr>
<td>Slc3a2</td>
<td>MGI:96955</td>
<td>solute carrier family 3 (activators of dibasic and neutral amino acid</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transport), member 2</td>
</tr>
<tr>
<td>Slc7a3</td>
<td>MGI:1100521</td>
<td>solute carrier family 7 (cationic amino acid transporter, y+ system),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>member 3</td>
</tr>
<tr>
<td>Slc7a7</td>
<td>MGI:1337120</td>
<td>solute carrier family 7 (cationic amino acid transporter, y+ system),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>member 7</td>
</tr>
<tr>
<td>Smad9</td>
<td>MGI:1859993</td>
<td>SMAD family member 9</td>
</tr>
<tr>
<td>Socs2</td>
<td>MGI:1201787</td>
<td>suppressor of cytokine signaling 2</td>
</tr>
<tr>
<td>Sox2</td>
<td>MGI:98364</td>
<td>SRY-box containing gene 2</td>
</tr>
<tr>
<td>Stat3</td>
<td>MGI:103038</td>
<td>signal transducer and activator of transcription 3</td>
</tr>
<tr>
<td>Suz12</td>
<td>MGI:1261758</td>
<td>suppressor of zeste 12 homolog (Drosophila)</td>
</tr>
<tr>
<td>Tcea3</td>
<td>MGI:1196908</td>
<td>transcription elongation factor A (SII), 3</td>
</tr>
<tr>
<td>Tdgf1</td>
<td>MGI:98658</td>
<td>teratocarcinoma-derived growth factor 1</td>
</tr>
<tr>
<td>Tdh</td>
<td>MGI:1926231</td>
<td>L-threonine dehydrogenase</td>
</tr>
<tr>
<td>Tfcp2l</td>
<td>MGI:2444691</td>
<td>transcription factor CP2-like 1</td>
</tr>
<tr>
<td>Trh</td>
<td>MGI:98823</td>
<td>thyrotropin releasing hormone</td>
</tr>
<tr>
<td>Trim28</td>
<td>MGI:109274</td>
<td>tripartite motif-containing 28</td>
</tr>
<tr>
<td>Trp53</td>
<td>MGI:98834</td>
<td>transformation related protein 53</td>
</tr>
<tr>
<td>Upp1</td>
<td>MGI:1097668</td>
<td>uridine phosphorylase 1</td>
</tr>
<tr>
<td>Wnt10b</td>
<td>MGI:108061</td>
<td>wingless related MMTV integration site 10b</td>
</tr>
<tr>
<td>Wnt11</td>
<td>MGI:101948</td>
<td>wingless-related MMTV integration site 11</td>
</tr>
<tr>
<td>Wnt16</td>
<td>MGI:2136018</td>
<td>wingless-related MMTV integration site 16</td>
</tr>
<tr>
<td>Gene</td>
<td>MGI</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>Wnt2</td>
<td>MGI:98954</td>
<td>wingless-related MMTV integration site 2</td>
</tr>
<tr>
<td>Wnt3a</td>
<td>MGI:98956</td>
<td>wingless-related MMTV integration site 3A</td>
</tr>
<tr>
<td>Wnt4</td>
<td>MGI:98957</td>
<td>wingless-related MMTV integration site 4</td>
</tr>
<tr>
<td>Wnt5a</td>
<td>MGI:98958</td>
<td>wingless-related MMTV integration site 5A</td>
</tr>
<tr>
<td>Yy1</td>
<td>MGI:99150</td>
<td>YY1 transcription factor</td>
</tr>
<tr>
<td>Zfp238</td>
<td>MGI:1353609</td>
<td>zinc finger protein 238</td>
</tr>
<tr>
<td>Zfp296</td>
<td>MGI:1926956</td>
<td>zinc finger protein 296</td>
</tr>
<tr>
<td>Zfp42</td>
<td>MGI:99187</td>
<td>zinc finger protein 42</td>
</tr>
<tr>
<td>Zfp428</td>
<td>MGI:1916463</td>
<td>zinc finger protein 428</td>
</tr>
<tr>
<td>Zswim1</td>
<td>MGI:1919221</td>
<td>zinc finger, SWIM domain containing 1</td>
</tr>
</tbody>
</table>
This appendix contains supplemental figures for Chapters 2 and 4. These figures are or will be available for download from the journal in which the work was published.
Figure B.1. mESC Network Performance Evaluation through Cross Validation. A. 4-Fold Gene Cross Validation. In addition to performance evaluations presented in Figure 2.3, we conducted 4-fold gene cross validation by removing 25 percent of genes in the gold standard training set. We first divided our list of 21,291 protein-coding genes into four sub-lists, then used these lists to create four sets of test (25 percent) and training (75 percent) files. Regularization reduced some overfitting; the remainder was largely corrected for through out-of-bag averaging. B. Leave-One-Gene-Out Cross Validation. We performed a variation on Jack Knife cross validation by removing edges containing a gene known to be essential for mESC self-renewal and pluripotency: Lif, Nanog, Pou5f1, and Sox2. Comparison of performance results before and after regularization illustrate the ability of the Bayes net to learn even when examples that included these key self-renewal genes were removed from the training set.
Figure B.2. Regularization and Self-Renewal Connectivity. **A, B.** Effect of Regularization on mESC and Test Networks. **A.** Self-Renewal Connectivity Strength of Gold Standard Genes. Regularization was performed to account for redundant information in large data compendiums. We adjusted regularization parameters to achieve a similar edge weight distribution pattern in both mESC-only and Mmu superset networks (see Material and Methods for details). **A.** Prior to regularization (yellow distributions), the mESC network had ~25 million edges with an inferred edge weight greater than 0.2, and a notable increase in edges near 1.0. After regularization (red distributions) to reduce the impact of redundant information, the number of edges with an inference score greater than 0.2 decreased to ~18 million. This process served to reduce the pile-up among edges with posterior edge weights at or near 1.0 caused by redundant evidence. **B.** The superset network before regularization had ~20 million edges with an inference score greater than 0.2 and a very sharp increase in edges near 1.0. After regularization, the superset network had only ~7.7 million edges with an inference score greater than 0.2. **C.** Comparisons of mESC and superset network positive gold standard gene correlations. The self-renewal correlation (SRC) score is a “guilt by association” measure of the strength of functional linkage between a given gene and most highly connected genes in the mESC positive gold standard, those with a gold standard functional correlation greater than 0.25. This SRC value (shown in green) is closely related to scaled degree (blue) and gold standard correlation (red), but gives a purer measure of correlation without noise introduced by incomplete biological knowledge. This distribution of the gold standard edges shows that some edges are more supported by the evidential data than others.
Figure B.3. Dataset Classes Supporting Top mESC Network Edges. High-confidence edges in the mESC network are supported by a diverse set of high-throughput data, illustrating that to achieve the best network performance in terms of biological relevance, data from many different types of experiments should be included in the data compendium. The Bayes net found Protein-DNA binding similarity profiles, followed by gene expression data, the most informative. Molecular interaction data downloaded from online resources and raw RNAi data were the least informative, largely due to the limited scope of these data. Supplemental Table S12 provides more details on the individual datasets that support the top 0.01% of network edges. A. The 639 top ranked mESC edges, those with a rank order of 1 and inferred edge weight ≥ 0.9999, were largely supported by transcription factor binding similarity profiles (dot products) calculated using protein-DNA binding data from ChIP-Seq and ChIP-Chip studies. Our Bayes net finds protein-DNA datasets the most reliable and dot products between vectors of individual transcription factor binding scores provide much more information than individual ChIP-based arrays. B. The top 0.01% of edges, 22,664 edges with an inferred weight ≥ 0.9954, were supported by a slightly different distribution of high-throughput data, with a smaller percentage of gene expression datasets contributing to edge weight. These top ranked edges are predominantly supported by dot products from evidential data from 12 ChIP-based studies with arrays for well-characterized genes known to play a key regulatory role in mESCs.
Figure B.4. Data Compendium Tests. A. To evaluate how compendium size contributed to the Bayes net’s ability to learn, we generated test compendiums of different sizes (from 5 to 55 randomly selected datasets) using the 58 mESC gene expression datasets included in this study. AUCs plateaued at approximately 0.65 at ~45 datasets or ~600-700 conditions. B. To compare the effect of using “inappropriate” non-cell-type-specific data compendiums of roughly the same size as the mESC network, we created three additional test compendiums using Mmu superset data: the top 60 gene expression datasets (as ranked by the Bayes net), the top 60 gene expression and PPI datasets, and 60 randomly selected datasets. We generated networks using these compendiums and the same mESC self-renewal gold standard and master gene list. All three Mmu test networks achieved higher AUCs than the mESC expression only network, but contained fewer high-confidence top ranked edges with a posterior edge weight ≥ 0.9999. Only the mESC network had edges with a posterior edge weight of 1 that involved genes that were highly enriched for biological processes associated with stem cell maintenance, embryonic development, and cell cycle controls (Supplemental Table S13). These results underscored that computation performance metrics alone are insufficient for evaluating cell-type-specific networks. C. Four-fold gold standard edge cross validation of these networks showed evidence of mild overfitting in all test networks. D. Correlation of gold standard edge weights showed the biological content of networks varied greatly. The mESC gene expression network was most similar to the Mmu Top 60 gene expression network ($R^2 = 0.36$), but there were stark differences. Many edges between developmental signaling pathway pairs, such as Fgfr3 – Araf and Bmpr1 – Mapk1, that were strongly supported in the Mmu Top 60 compendium, yet weakly supported in the mESC only network. Similarly edges involving canonical pluripotency factors, such as Pou5f1 – Cdkn1a and Nanog – Phc2, were strongly supported in the mESC network but not in the Mmu Top 60.
**A**

**mESC Data Compendium Tests**

- **Average Network AUC versus Number of Datasets**
  - Number of Datasets in the Integrated Data Compendium: 0.95

**mESC-Specific Data Compendium Tests**

- **Predictive Network AUC versus Number of Conditions**
  - Number of Conditions in Integrated Data Compendium: 0.95

**B**

<table>
<thead>
<tr>
<th>Test Network</th>
<th>Data Compendium Composition</th>
<th>AUC</th>
<th>Top Ranked Edges/Genes*</th>
</tr>
</thead>
<tbody>
<tr>
<td>mESC Exp68</td>
<td>58 mESC gene expression datasets</td>
<td>0.6858</td>
<td>22,123(1771)</td>
</tr>
<tr>
<td>Superset Top 50ex</td>
<td>Top 50 contributing Mmu gene expression datasets</td>
<td>0.7307</td>
<td>311,230</td>
</tr>
<tr>
<td>Superset Top50</td>
<td>Top 50 contributing Mmu datasets (gene expression &amp; PPI)</td>
<td>0.7381</td>
<td>463,323</td>
</tr>
<tr>
<td>Superset Rand50</td>
<td>50 randomly selected Mmu gene expression datasets</td>
<td>0.6981</td>
<td>112</td>
</tr>
</tbody>
</table>

*number of edges with a posterior edge weight > 0.01 and the number of unique genes involved in those edges.

**C**

**mESC Exp68 Network 4-fold Cross Validation**

- Receiver Operating Characteristic (ROC) Curve
  - Average Training AUC: 0.7136
  - Average Test AUC: 0.6895

**Superset Top50exp Network 4-fold Cross Validation**

- Receiver Operating Characteristic (ROC) Curve
  - Average Training AUC: 0.7136
  - Average Test AUC: 0.6895

**Superset Top50 Network 4-fold Cross Validation**

- Receiver Operating Characteristic (ROC) Curve
  - Average Training AUC: 0.7124
  - Average Test AUC: 0.6746

**Superset Rand50 Network 4-fold Cross Validation**

- Receiver Operating Characteristic (ROC) Curve
  - Average Training AUC: 0.7102
  - Average Test AUC: 0.6722

**D**

**Positive Gold Standard Posterior**

- Superset Top50 versus mESC Exp68
  - R^2 = 0.8014

**Positive Gold Standard Posterior**

- Superset Top50 versus mESC Exp58 Network
  - R^2 = 0.8014

**Positive Gold Standard Posterior**

- Superset Rand50 versus mESC Exp58 Network
  - R^2 = 0.7886
Figure B.5. Datasets Supporting Tdh Connectivity to Gold Standard Genes. The top 10 datasets supporting high-confidence edges between Tdh and gold standard genes Sox2, Nanog, Nrb01, and Rif1 are predominantly composed of transcription factor binding similarity profiles, suggesting that Tdh is likely to be a target of genes involved in the regulatory circuitry of pluripotency and self-renewal. This is consistent with recent evidence that mESCs have unusual, accelerated cell-cycle kinetics to support rapid growth phase of early embryonic development that require high amounts of ATP as fuel.
Figure B.6. Network Performance Evaluations. A. Computational assessment of network performance using standard machine learning metrics showed that the area under the Receiver Operating Characteristic (ROC) curve (AUC) for the mESC network was 0.7514; 0.7289 after regularization and out of bag averaging, the AUC was 0.7165. AUCs for the hESC network were 0.7802 before and 0.7520 after regularization. B. We conducted 4-fold network cross validations by removing 25% of edges in the gold standard (4-fold Gold Standard). ROC curves showed a small amount of overfitting, most apparent in cross validations for which we removed 25% of genes (rather than edges) from the network training set.
These supplemental notes contain step-by-step instructions for preprocessing microarray gene expression data to use as evidential data and for generating cell-type-specific naïve Bayesian networks using C++ tools from the Sleipnir Library of Computational Functional Genomics (www.huttenhower.org/sleipnir/index.html). The ruby scripts mentioned in this documentation are available on request from the Hibbs Laboratory StemSight team (email stemsight@jax.org).

**Gene Expression Data Pipeline**

**Initial Data Collection**

This will vary from source to source, but the output of the first phase is always a preclustered (PCL) formatted file and meta information about the data.

**For Gene Expression Omnibus (GEO) Dataset Files**

To collect curated GEO dataset (GDS) files and prepare them for processing using Sleipnir tools, perform the following steps:

**Download All GDS Datasets From GEO.** By anonymous logon to the NCBI ftp site, all available microarray data can be downloaded from ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SOFT/GDS/ as
gzipped soft formatted files. Currently the easiest way to get all data for a single organism is to
download all of the available data and verify organism after reading out header information.

Collect Information From All Datasets and Convert to PCL Format. We can identify information
about each dataset, such as organism, platform used, sample descriptions, etc. from the header
information in the soft files. We’ll do this at the same time as we read the files to convert them
to pcl format. To do this, run the ruby script allSoft2Pcl.rb and store the output:

    ruby allSoft2Pcl.rb <path_to_soft_files> > <info_file>

This will create an info file that contains basic information about each file, as well as
creating a pcl formatted file for every *.soft file in the specified directory. Alternately, files can
be converted one at a time using the script convertSoft2Pcl.rb.

    In a properly formatted pcl file, there are two header lines, the first two columns are
    probe IDs or identifiers, followed by a weight column, then values for specific arrays. It should
    be a tab-delimited files that looks like this:

    | YORF | NAME | GWEIGHT | <Array1> | <Array2> | <Array3> |
    |------|------|---------|---------|---------|---------|
    | EWEIGHT | 1    | 1       | 1       | 1       | 1       |
    | <Gene1> | <Gene1> | 1     | <value1.1> | <value2.1> | <value3.1> |
    | <Gene2> | <Gene2> | 1     | <value1.2> | <value2.2> | <value3.2> |
    | <Gene3> | <Gene3> | 1     | <value1.3> | <value2.3> | <value3.3> |

For GEO Series Files

    GEO Series (GSE) Files have been uploaded to the GEO repository by researchers, but
have not by curated by through GEO. These are the source files for curated GDS datasets, so be
careful you don’t use both the GDS and GSE for the same study. In some cases, there is a
problem with the uploaded GSE files that prevent it from passing GEO curation and there is no corresponding GDS file.

Perform Custom Download of Specific GSE Files of Interest. This requires more customized searches on the NCBI GEO site (www.ncbi.nlm.nih.gov/geo/) to identify datasets that contain microarray data of interest. Identify key words that will retrieve all potential GSE files of interest.

Manually Curate GSE Files to Prepare for Pipeline Processing. Using a text editor, such as Gnu EMACS or VI, remove any arrays that are not specific to your study. If more than one platform is used, split the series file into two or more files, one for each platform

Convert GSE Files into PCL Format. GSE files are all “special” in that they all have (usually minor) problems with the file format or content that causes them not to be automatically curated into GDS files. For that reason, you will need to do some extra work to prepare the GSE family.soft files for processing in our pipeline. Use the seriesFamilyParser.rb script:

    ruby seriesFamilyParser.rb <filename> > <filename.pcl>

This script will prompt you to answer a series of questions about the microarray data to identify which column contains the probeID, gene name (symbol), values, whether the data has one or two channels, if it has been log transformed, etc. It will then generate an output file in pcl format

Generate Info File for All GSE PCL Files to be Processed in Pipeline. Ensure all pcl files generated by the seriesFamilyParser script are in the same directory. Create an InfoFile that contains general reference information about all pcl files that will be used in subsequent pipeline processing steps.
To do this, run the AllPcl2Info.rb script:

```
ruby AllPcl2Info.rb <path to pcl files> > <info_file>
```

Other data repositories and sources used by the preprocessing steps vary. Any files to be used as evidential data must be in a PCL matrix format.

**Data Normalization Pipeline**

**Insert Missing Values**

While non-numeric entries will have already been turned into missing values, some laboratories use values that are better characterized as missing. Run the script insertMissingValues.rb to address these:

```
ruby insertMissingValues.rb <pcl_file> <info_file>
```

This script will append .mv to the pcl file as the newly created output.

**Impute Missing Values and Remove Probes with Few Values Present**

Since later methods require a full data matrix, we use KNN-Impute to fill in missing values with their most likely values. However, this process is unreliable when few actual values are present, so probes are required to be present in at least 70% of the conditions to be kept. Run the script runKnnImpute.rb to do this:

```
ruby runKnnImpute.rb <pcl_file> <info_file> <path_to_KNNImputer>
```

This script will run KNNImputer with the proper parameters and pre/post-processing, and append .knn to the pcl file as the newly created output. KNNImputer is part of the Sleipnir library (Huttenhower et al., 2008).
Map Probes to Genes

Using the script `mapGeneNames.rb`, potential aliases and conflicts are resolved by translating all gene names to a common standard. This script requires a file that includes the organisms of interest and points to alias mapping files. These files should contain 2 tab-delimited columns, the 1st of which contains single aliases, and the 2nd of which contains standard names that are | (bar)-delimited. This script can be run with:

```
ruby mapGeneNames.rb <pcl_file> <info_file> <org_file>
```

This will append .map to the newly created output, if the organism of the file is included in the script hard-coded list. Otherwise, no new output file is made.

Average Together Duplicate Genes

The Java program `MeanGenesThatAgree.jar` will calculate the mean of all probes that meet a maximum likelihood test of measuring the same gene. Specifically, this method compares the distribution of Euclidean distances between pairs of probes annotated to the same gene to the distance distribution between randomly selected probe pairs. Probes are averaged together and annotated to a single gene if their distance is more likely to be drawn from the annotated probe distribution than from the random distribution. This script can be run with:

```
java -Xmx2g -jar MeanGenesThatAgree.jar <pcl_file> 1 > <out_file>
```

Name the outfile to append .avg to the newly created output file.
Final Numeric Clean Up and Consolidation

Some data files require a final numerical transformation (logarithm of all measurements, performed with a call to DivLogNorm.jar). This script performs this if necessary, and always moves the final file to an organism specific directory:

```
ruby collectFinalData.rb <pcl_file> <info_file> <script_path> <output_path>
```

**Note:** There is also a script that runs the entire pipeline for a single input pcl. It can be run with:

```
ruby fullyNormalizeDataset.rb <pcl_file> <info_file> <org_file> <path_to_pipeline_scripts> <path_to_KNNImputer> <output_path> [REMAKE]
```

---

**Using Sleipnir Tools to Generate Cell-Type-Specific Predictive Networks**

These are the basic steps for using Sleipnir tools to generate cell-type-specific predictive networks. You will need to install the Sleipnir library of tools on your server before you can begin. The computational resources and time required to use these tools varies depending on the size of your data compendium, gold standard, and master gene list. On the Hibbs Lab cluster (~250 compute cores), the mESC network, with 164 evidential datasets, takes ~4 hours to compute. In contrast, the superset network, with 810 datasets, takes ~16 hours. Cross validation and bagging is more time-consuming and the total time required depends on the number of bootstrap runs required. In general, it takes about 2 weeks to perform regularization and bootstrap aggregation on a final network and to prepare the supporting files required to publish the network online in our visualization tool.
What you need

- Gold Standard positive and negative lists of gene pairs (all genes should be referenced using consistent IDs, e.g. MGI:ID, Entrezgene ID, official gene symbol)
- Two gene lists of all protein-coding genes in your genome of interest, one numbered.
- Optional: Context gene list(s) of all genes within a biological context of interest
- Optional: Context file that lists all context gene lists to be used
- Datasets in .pcl format
- Sleipnir tools: Distancer, Dat2Dab, Counter, DChecker, BNs2Txt, and M1er. For installation instructions, go to libsleipnir.bitbucket.org/
- R Statistical Software
- Microsoft Excel

Prepare Network Directories and Input Files

To generate naïve Bayesian Networks using integrated genomic data, you will need a gold standard training set and a list of protein-coding genes and directory containing evidential data files. Because there are several steps that generate intermediate files, create a directory structure to organize these files for each network, such as:

<table>
<thead>
<tr>
<th>Directory Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td>Network folder</td>
</tr>
<tr>
<td>dabs</td>
<td>Integrated cell-type-specific data .dabs and .quants file folder</td>
</tr>
<tr>
<td>gl</td>
<td>Gene list folder</td>
</tr>
<tr>
<td>gs</td>
<td>Gold standard folder</td>
</tr>
<tr>
<td>counts</td>
<td>Conditional probability table counts folder</td>
</tr>
<tr>
<td>results</td>
<td>Prediction results folder</td>
</tr>
<tr>
<td>xval</td>
<td>Cross validation folder</td>
</tr>
<tr>
<td>analysis</td>
<td>Network analysis folder</td>
</tr>
</tbody>
</table>

Create Gold Standard Answer File. Screen your positive and negative gene pair lists to ensure there are no duplicate gene pairs in these lists (i.e. no positive pairs should appear in the negative list). Combine your positive and negative gene lists to create one tab-delimited text file, with three columns. The format should be:

<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>Answer File Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive gene pair</td>
<td>MGI</td>
</tr>
<tr>
<td>negative gene pair</td>
<td>MGI</td>
</tr>
</tbody>
</table>
Once you have an answer file that contains positive and negative gene pairs and their respective relationship values, convert the file to .dab format, which is a compressed file format that can be read and processed by Sleipnir tools. Use Dat2Dab to convert a text file with Unix (lf) line breaks to .dab format. Depending on the directory path to your installed Sleipnir Tools, the basic command is:

```
Dat2Dab –i <answers.dat> –o <answers.dab>
```

Where <answers.dat> is the name of your input file and <answers.dab> is your output file. Use this same tool to convert .dab files to readable text format.

Even though your answer file should have only two classes (0,1), you need to create a file that defines answer file bins for the Bayes net. To do this, create a corresponding quantification (.quant) file for every .dab file you generate. This is simply a tab-delimited text file that contains one line that specifies the bins that will be used to discretize continuous data. For the answer file, which has only two values (0 and 1), there are only two bins:

```
0 <tab> 1
```

You can create this file in any text editor. The file should have the same name as the .dab file, with a .quant extension. For example:

```
Answer File:   gs_answers.dab
Answer Quant File: gs_answers.quant
```

Save these two files (.dab and .quant) to your working Bayesnet directory. For more information on Dat2Dab options, see libisleipnir.bitbucket.org/Dat2Dab.html.
Convert Dataset PCL Files to .dab Format. Run Distancer to convert pcl files to .dab format that can be processed by Sleipnir. These .dab files contain pairwise similarity score calculated from the microarray data in the .pcl files. The basic command is:

```
Distancer -i <dataset1.pcl> -o <dataset1.dab>
```

Where dataset1.pcl is the name of the input file you want to convert and data.dab is the name of the output file you’ll use when generating Bayes nets with Sleipnir tools. Save these evidential data .dab files to a dabs/ subdirectory within your working network directory. If you have multiple data compendiums, create a separate subdirectory for each (for example: dabs_mESC/, dabs_superset/, dabs_minset/).

**Note:** The default Distancer command produces a z-score (z-transformed Pearson correlations) for each gene pair in the .pcl file. If you want to test other similarity measures, add a -d switch to the command line. Valid Distancer -d values are: "pearson", "euclidean", "kendalls", "kolm-smir", "spearman", "pearnorm", "hypergeom", "innerprod", "bininnerprod", "quickpear", "mutinfo", "relauc", "pearsig". The default is "pearnorm".

For more information on Distancer options, see libsleipnir.bitbucket.org/Distancer.html.

Verify .dab Files Were Converted and Transformed Properly. Use R to check the normalization curves of the data in the .dab files created by Distancer. The .dab files are not in a format R can read, so you must first convert them back to tab-delimited data files (.dat files).

Run Dat2Dab to create a set of .dat files from your newly created .dab files. The basic Dat2Dab command is:

```
Dat2Dab -i <dataset1.dab> -o <dataset1.dat>
```

Where dataset1.dab is the name of your machine-readable input file and dataset1.dat is the text outfile. If you used the default Pearson Normal Distancer option to produce evidential
data .dab files, use R or your statistics program of choice to produce a histogram from the .dat file and check the distribution.

Verify the normalization curve looks good. If so, move the .dab file to a dabs/subdirectory in your working directory. If the normalization curve don’t look “normal,” refer to the source dataset in your master file directory or repository and double-check the number of samples in the source datasets. Microarray datasets must have at least 4 samples to be processed correctly.

Note that .dat files generated from .dab files may be quite large. You may want to create a script that converts .dabs to dats, produces the R histogram, then deletes the intermediate .dats if disk space is limited.

Create Corresponding Quant Files for Each Dataset .dab File. Expression values in high-throughput microarray data files are continuous. To discretize this data for the Bayesian classifier, you need to create quant files for each dataset in your <input_file> subdirectory. In our case, the quant file for all microarray datasets was a one-line, tab-delimited text file containing these seven bins:

-1.5 -0.5 0.5 1.5 2.5 3.5 4.5

Verify that you have one .quant file for every .dab file in your subdirectory and that all files use a consistent naming format. For example:

transfac.dab
transfac.quant
GSE9978.dab
GSE9978.quant

Double-Check Working Directories and Files. Verify that you have created dabs, counts, and results subdirectories for your network files, and that you have all the files you need to generate a Bayes net: your answer files (.dab and quant), two gene lists (one numbered), a optional
context file (a subset of protein-coding genes, such as for a biological process or developmental
stage), evidential dataset files in .dab format (in a separate dabs/subdirectory with
the corresponding quant files).

Note: Most issues with generating Bayes nets using the Sleipnir Counter tool are due to
missing files, misnamed files, errors in paths, etc. This checkpoint will help you avoid potential
problems and unnecessary angst.

Create Networks

Generate Conditional Probability Table (CPT) Counts (Counter –w Step). Use Sleipnir Counter to
generate a counts file in your output directory. This counts file summarizes the number of data
values in the each discretized dataset.dab file relative to the functional gold standard in the
answer file. To generate global counts, use the following Counter command:

    Counter –w <answers.dab> -d <dabs/> - o <counts/>

Where answers.dab is the name of your answers file, dabs/ is the name of the directory
that contains your dataset .dab and quant files, and counts/ is the name of the output directory.
This will produce a global.txt counts file in the /counts directory.

Note: If you are generating Bayes nets with multiple context-specific gene lists (such as
for developmental stages or tissue types), you’ll need to run this step for each context. To
generate counts for each context, add the context file name to the end of the command line:

    Counter –w <answers.dab> -d <input_data/> - o <counts/> <context_gene_list.txt>

This will produce a context counts file (with the same name as the context gene list
file) in the output directory.
Open the global.txt counts file to verify it looks correct. These files should contain counts of positive and negative values in each bin for each dataset in the input directory.

Here’s an example of the first several rows of a global counts file that references the total number of datasets, the number of negative and positive genes, and counts for each dataset (only 3 of a total of 65 dataset counts are shown):

```
global  65
  18239  1709
dataset1
   448   1870  3230  1723   409   69   16
   114    352   556   352   104   20    8
dataset2
   261   1041  1773   987   209   46   15
    74    244   409   246   65   16   11
dataset3
   118    496   773   531   96   16    5
   65    254   372   229   59    8    2
```

Note that if you used a global gene list for your context file, your global and context count files should be the same. For more information on Counter, see libsleipnir.bitbucket.org/Counter.html.

**Produce a Dataset List.** Use the global counts file to create a numbered list of datasets in the order in which Counter processed them. This should be a tab-delimited text file that contains the sequence number and the data file name (with no extension). For example,

```
1 <tab> dataset1
2 <tab> dataset2
3 <tab> dataset3
4 <tab> dataset4
```

Save this dataset file to your network directory. You may find it easier to write a script to create a numbered dataset file using the global.txt output file from the first Counter step. (If you are using multiple contexts, you will need to create a context.con file that contains a
numbered list of all context files to be used when generating the Bayes net. A context.con file is a tab-delimited text file that contains one line in the following format:

1  <tab>  context1_gene_list  <tab>  context1_gene_list
2  <tab>  context2_gene_list  <tab>  context2_gene_list
3  <tab>  context3_gene_list  <tab>  context3_gene_list

Save the context.con text file to your gene list subdirectory.

**Generate CPTs (Counter –k Step).** Use **Counter** to populate CPTs that comprise the naïve Bayesian network using the counts files and the dataset list created above. This step will produce a networks.bin output file that contains the Bayes net CPTs.

To generate CPTs, use the **Counter** -k command:

```
Counter -k <counts/> -o <networks.bin> -s <datasets.txt> -b <counts/global.txt>
```

Where counts/ is the name of the output directory you specified in the **Counter** –w step, networks.bin is the name of the Bayes Net output file, counts/global.txt is the path and file name of the global counts file created the –w step.

The **Counter** –k command will produce a networks.bin output file that contains CPTs.

**Optional Regularization Parameter**

If you are performing regularization, append a pseudocount parameter to the **Counter** –k command:

```
-p <pseudocount>
```

Where <pseudocount> is the regularization parameter used to modulate the strength of regularization required as implied by the strength of the prior (higher pseudocount values weaken the influence of redundant datasets.) You will need to test several pseudocount values (e.g., 1, 5, 10, 20, 30, ...) to determine the optimal regularization strength for your data compendium.
Optional Context Parameter. If you are using contexts, append a context parameter to the Counter –k command:

- X <context.con>

Where <context.con> is the text file that contains the lists of context files. (Note that the context file switch is a capital X.)

Make Predictions (Counter –n Step). The Counter –n command performs Bayesian inference for the classifier saved in the networks.bin file created as output of the Counter –K step. It produces a predictive functional relationship network graph file. If no context files are specified, it will generate a single global.dab file containing edges for all possible gene pairs and inference scores (edge weight) for each edge.

To infer functional relationships based on the CPT tables generated in the –k step, use the Counter –n command:

Counter –n <networks.bin> -o <results/> -d <dabs/> -s <datasets.txt> -e <gene_list_numbered.txt>

Where networks.bin is the name of the Bayes Net output file created in the –k step, results/ is the name of the output directory, dabs/ is the name of the data directory that contains .dab and quant files, datasets.txt is the name of the dataset list, and gene_list.txt is a numbered list of all protein-coding genes. If you are using one or more contexts files, append them to the end of this command line.

If you are using contexts, append the name of your context file to this command (after the –e switch) and Counter will create a .dab file for each context.

Additional Option for Regularization. If you are regularizing your data, you will need to append a mutual information alpha file parameter to the Counter –n command:
-a <dataset_weights.dat>

Where <dataset_weights.dat> is the alpha file that summarizes the mutual information scores for each dataset in the data compendium. To generate this alpha file, use Mler to create a mutual information (MI) file for each dataset in your dabs/ directory, create an MI matrix using Combine_to_matrix.rb, and scale the raw scores with half2weights.rb. The basic commands are:

Mler -y 0 <dataset.dab> dabs/*.dab > dataset.dab.mi
Ruby combine_to_matrix <path_to_MI_files>
ruby half2weights.rb <bn_xdsl file) < <matrix.dat> > outfile

The Ruby scripts referenced are available on the Sleipnir site. To generate the xdsl file, convert the networks.bin file to an xdsl file using Sleipnir Bns2Text.

**Analyze Bayes Net Performance**

Use DChecker to produce information needed to generate statistics required to analyze performance. To run DChecker, use the command:

DChecker -w <answer.dab> -i <results/global.dab> -g <gene_list.txt> > <performance_results.txt>

Where <answers.dab> is the gold standard answer file specified in the –w step, results/global is the path and file name of the network output file generated in the –n step, gene_list.txt is a list of all protein-coding genes, and performance_results.txt is the name of the output file that will capture DChecker information. This will produce counts of True Positives, True Negatives, False Positives, and False Negatives for 1000 cuts and an AUC score (for the Area Under Curve). Version 3.0 of Sleipnir DChecker also includes columns for Precision (PR) and Recall (RC). If you use these calculated values, note that the column labels are transposed (as of the time of publication). For finer control over analysis, you can use a –b switch to specify
the exact number of bins for quantile sorting by increasing the number of cuts from 1000 to 10,000 or 20,000, depending on the total number of genes in your gene list and the amount of detail you want in your performance analysis. For example if you have ~20,000 genes in your master gene list, and want to specify one bin per gene, use this command:

DChecker –w <answers.dab> -i <results/global.dab> -g <gene_list.txt> -b <20000> > <performance_results.txt>

Use the output file counts to calculate performance measurements of Precision, Recall, Sensitivity or True Positive Rate (TPR), and 1-Specificity or False Positive Rate (FPR):

<table>
<thead>
<tr>
<th>Metric</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>TP/(TP+FP)</td>
</tr>
<tr>
<td>Recall</td>
<td>TP/(TP+FN)</td>
</tr>
<tr>
<td>TPR</td>
<td>TP/(TP+FN)</td>
</tr>
<tr>
<td>FPR</td>
<td>FP/(FP+TN)</td>
</tr>
</tbody>
</table>

Use R or Excel to produce a Receiver Operating Characteristic (ROC) curve, which is a graphical plot of TPR versus FPR. The AUC score reported in the final row of the DChecker outfile is the area under the ROC curve. You should also produce a plot of Precision versus Recall (use logarithmic scale for Recall on the X-axis), and other performance metrics as warranted.

The ROC and PR curves are only two performance metrics you can use to analyze network performance, but they are a good starting point. If you have questions, please contact the StemSight team at stemsight@jax.org. Good luck with your networks!

For more information about DChecker options, see libslipnir.bitbucket.org/DChecker.html.

For more information about performance statistics for evaluating naïve Bayes Nets, see en.wikipedia.org/wiki/Receiver_operating_characteristic.

For more information about StemSight Bayes Nets from the Hibbs Lab, see www.stemsight.org.
These Supplemental Notes contain detailed materials and methods on how to generate species-specific probabilistic networks using a Bayesian network machine learning framework for genomic data integration. For more information about the tools and methods described here, see our previous publication [1].

Collection and Preparation of Training and Evidential Stem Cell Knowledge

Supervised Bayesian network machine learning requires a consistently integrated collection of diverse high-throughput evidential datasets, coupled with a reliable reference gold standard (prior knowledge) for training and evaluation. To ensure high quality, consistent, and comprehensive system input, we developed a rigorous protocol for gathering and preprocessing input datasets, and carefully documented our methods for developing tailored gold standards.

Preparation of Evidential High-Throughput mESC and hESC Dataset Compendiums

We preprocessed, normalized, and standardized a comprehensive set of mESC input data from high-throughput experiments using microarrays, ChIP-Chip, ChIP-Seq, affinity purification followed by mass spectrometry (AP-MS), and whole genome small interfering RNA (siRNA) screens, plus molecular interaction and phylogenetic data not specific to any cell type. A complete list of data sources used is provided in Supplemental Tables S4 (mESC data) and S5 (hESC data). This data was mapped to MGI Gene IDs, and preprocessed into ~6 billion pairwise values used as features for classification. Through this process, each protein-coding gene pair
(in each dataset) was assigned a similarity score, based on Euclidean or Pearson correlation distance measures between genes as previously described [1]. Pearson correlations were normalized using Fisher’s Z-transform, shifted by the mean, and divided by the dataset standard deviation to yield a collection of pairwise similarity scores with an approximately normal distribution ~N(0,1). Values were binned into discrete ranges for use as classification features in our Bayes net integration (details as follow).

Microarray Expression Data. Microarray gene expression data files were downloaded from the Gene Expression Omnibus (GEO) [105] and processed as previously described [1]. Raw microarray datasets available in Affymetrix CEL format were normalized using the Robust Multichip Average (RMA) function in Bioconductor R/affy package (version 2.5, R version 2.10.1). Brainarray ENTREZG custom chip definition files (CDFs), which reflect the most recent gene and probe sequences, were used to map probes to genes (version 12.1.0) [62, 106, 107]. Microarray data downloaded from public databases was preprocessed as previously described [1]. The resulting standardized datasets were then converted to a PreCLustered (PCL) format and distilled into a set of pairwise similarity scores using Pearson correlation followed by Fisher’s z transformation to measure the strength of the linear relationship between gene expression values for all possible gene pairs in the study (Equations D.1, D.2). If the dataset contained only two samples (conditions), we used Euclidian distance (Equation D.3) to measure the strength of relationship.

\[
r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}
\]

(D.1)

\[
z = \frac{1}{2} \ln \frac{1 + r}{1 - r}
\]

(D.2)
Where $r$ is the Pearson correlation coefficient calculated from the microarray profiles, and $z$ is the Fisher Z-transformed correlation.

$$L_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$  \hspace{1cm} \text{(D.3)}

Where $L_2$ is the Euclidean distance function calculated between measurements for genes $x$ and $y$.

**Chromatin Immunoprecipitation (ChIP) followed by Microarray (ChIP-Chip) Data.** Raw high-throughput ChIP data was obtained from online supplemental materials and from contributing author websites. We organized data into a consistent ChIP data matrix format, mapping bait and target gene IDs to systematic MGI gene IDs. Data was processed into a pairwise format in two ways: first, we generated separate data files for each transcription factor used in the study (e.g. input pairs connecting each transcription factor with its putative targets as inputs or features); and second, we generated transcription-factor-binding similarity profiles between all gene pairs (e.g. values determined by the number of transcription factors shared by each gene pair). These similarity profiles were created by calculating dot products between vectors of individual transcription factor binding scores for each gene pair in the study (Equation 3).

$$D_{a,b} = \sum_{i=1}^{n} (a_i b_i) = (a_1 b_1) + (a_2 b_2) + \ldots + (a_n b_n)$$  \hspace{1cm} \text{(D.4)}

Where $D_{a,b}$ is the dot product score for the pair of genes $a$ and $b$, $n$ is the total number of transcription factors interrogated in the study, and $a_i$ and $b_i$ are binding scores (often binary) for genes $a$ and $b$ and the $i^{th}$ transcription factor.
ChIP Followed by High-Throughput Sequencing (ChIP-Seq) Data. Transcription factor binding site and gene association scores [89], based on the genomic location of the binding site closest to the transcription start site of expressed genes, were used as raw data and processed in the same manner as ChIP-Chip data. Broad peaks in ChIP-Seq data supplied in Browser Extensible Data (BED) format were mapped to gene regions using an NBCI annotation file for the genome build specified in the publication. We took the average, maximum, and sum of all peaks within a gene region for each sample, then created a matrix of gene hits (rows) by sample (column) and used Pearson correlation (Equation 1 and 2) or, if less than three samples were available, Euclidean distance (Equation 3), and dot product (Equation 4) similarity profiles to distill this matrix into pairwise gene associations. We used the Bayes net to evaluate which summary method (average, maximum or sum) was most relevant and selected the final datasets based on which best-supported edges in the positive gold standard.

RNA-Seq Data. Metafiles of preprocessed and annotated RNA-Seq reads were downloaded from GEO and preprocessed in the same manner as standardized microarray datasets. Raw data supplied in BED format were processed using the same method as ChIP-Seq data.

Methylation Profiling Data. The location of DNA methylation marks identified through high-throughput methylation profiling assays was mapped and preprocessed in the same way as ChIP-Seq data supplied in BED format.

Whole Genome Small Interfering RNA (siRNA) Screen Data. Raw siRNA data from primary screens represented the percent of differentiating cells upon exposure to siRNA knockdown of mESC self-renewal genes. We organized these values into a matrix format analogous to that used to preprocess microarray and ChIP data. A Euclidean distance measure was used to distill this data into pairwise similarity scores (Equation D.3).
Prepare Curated, Tissue-Specific Training Gold Standard

We developed a comprehensive positive gold standard for mouse and human ESC genes and gene products involved in self-renewal, curated from a literature library of 108 recent articles related to ESC self-renewal, pluripotency, and cell fate determination (Supplemental Table S2). This library was supplemented with developmental signaling pathway information from KEGG [49]. A list of all publications referenced in the ESC gold standard is available in Table S3. From this ESC reference, we extracted a list of 2077 “positive pairs” of gene or gene products experimentally validated to be functionally related in the context of mESC fate. To generate a negative gold standard, we first developed a master list of 17,342 protein-coding gene orthologs (Supplemental Table S1) derived from an MGI Human and Mouse Orthology Report (HMD_Human3.rpt for Build 37, downloaded December 05, 2012). We selected MGI IDs with a marker type of protein-coding genes, then culled out the 365 genes in the positive gold standard to create a list of 16,977 negative gold standard genes, from which we randomly generate a list of 20,770 “negative pairs”. The resulting gold standard answer file (Supplemental Table S10), consisting of a total of 22,847 gene pairs with a class distribution of 1:10 (positive:negative), was used as the prior knowledge to train the Bayes net.

Construct Bayesian Network and Infer Posterior Functional Relationship Scores

To perform naïve Bayesian network machine learning techniques, we computed the posterior probability of a functional relationship between gold standard gene/protein pairs given all evidential data [55, 60, 62, 110]. We used the Sleipnir library of C++ tools for machine
learning over genomic data [110] and the Structural Modeling, Inference, and Learning Engine (SMILE) C++ library, developed at the University of Pittsburgh [111] as previously described [1].

**Bayesian Network Training and Inference**

Conditional probability tables (CPTs) for each dataset were learned by counting the observed values in each dataset’s discretized bins for unrelated and related training gene pairs [51, 62]. Once learned, these CPTs were used to infer posterior functional relationship scores between pairs of genes or gene products. The posterior probability that two protein-coding genes participate in a self-renewal related biological process, given existing data, was calculated based on the prior probability of a functional relationship between genes and the conditional probability of observing evidential data given functional relationship status (Equation D.5) [51, 55].

\[
P(FR | E_1, E_2, ... E_n) = \frac{1}{Z} P(FR) \prod_{i=1}^{n} P(E_i | FR) \tag{D.5}
\]

Where \( FR \) is a hidden variable representing whether a gene pair is functionally related, \( P(FR=1) \) is the predicted probability that a pair is functionally related, \( E_i \) represents the evidence score of the gene pair for the \( i^{th} \) dataset, and \( Z \) is a normalization factor.

**Minimization of Network Overfitting**

We performed four-fold cross-validation (on both edges and gene in the gold standard) to determine classifier performance and generality as previously described [1]. To minimize overfitting, we did 20 bootstrap runs and performed bootstrap aggregation (i.e. bagging) as previously described [1].
Regularization

Bayes nets impose a strict assumption of independence between input data that is likely violated by many of our input datasets. This limitation can be largely mitigated through regularization of parameters to down weight the contribution of datasets with redundant information (Supplemental Figure B.6). Parameter regularization was performed using mutual information between datasets to weight the strength of prior belief for each dataset [62, 81]. Because the same subset of information could be shared many times among cell-type-specific data collected for the same set of genes, using the same array platform or under the same conditions (e.g., knockout, knockdown, overexpression, or constitutive expression), this regularization provided a quantitative estimate of the amount of redundant information contained in each dataset as compared to all other datasets in the compendium. We calculated a heuristic sum of mutual information relative to the Shannon entropy of each dataset [112] (Equation D.6), exponentially decreasing the weight of a dataset as the amount of shared information increased and incorporated these values into the formula for calculating posterior probability (Equation D.7) as previously described [51, 55].

\[
S_k = 1 + H(D_k)^{-1} \sum_{i \neq k} I(D_i; D_k)
\]

\[
P(FR_{i,j} | E_1, E_2, ..., E_n) = \frac{1}{Z} \prod_{k=1}^{n} \frac{\alpha P[D_k = d_k(g_1, g_2)] + \alpha^{S_k} - 1}{\alpha + |D_k| \alpha^{S_k-1}}
\]

Where \(S_k\) is a heuristic sum of shared information relative to the dataset’s entropy used to weight the strength of prior belief in a uniform distribution for the dataset, \(H\) refers to Shannon entropy, and \(I(D_i; D_j)\) refers to mutual information. Equation 7 is an variation on Equation 5, such that \(P(FR_{ij} | E_1, E_2, ..., E_n)\) is the predicted probability that there is a functional relationship between genes \(i\) and \(j\) given evidence in datasets \(1\) through \(n\), \(Z\) is a normalization
factor, $\alpha$ is a pseudocount regularization parameter used to modulate the strength of regularization required as implied by the strength of the prior (higher pseudocount values weaken influence of redundant datasets), and $D_k$ is the number of bins used to discretize continuous data values in dataset $K$. A low $S_k$ indicated the information contained in the dataset is highly unique, while a high score indicated the datasets contained shared (redundant) information. The redundancy score for each hESC dataset is listed in Supplemental Table P.5; we used a pseudocount value of 100 to regularize these hESC evidential data. The redundancy score for each mESC dataset used to train the Bayesian classifier is listed in Supplemental Table P.6. We conducted performance tests to evaluate effects of regularization on similarity score distributions in each evidential dataset and classifier performance, and selected an optimal pseudocount value of 70 to regularize mESC evidential data.

**Computationally Test and Validate Results**

We validated the accuracy of predicted functional relationships computationally using standard machine learning metrics and accepted protocols.

**Evaluation Metrics**

To assess network predictive accuracy, we used standard statistical performance measures for binary (true/false) classification tests: Receiver Operating Characteristic (ROC) Curves, Area Under the ROC Curve (AUC), Precision-Recall Curves (PRC), and Area Under the PRC (AUPRC) [51, 52]. A ROC curve is a two-dimensional graph of true positive rate (TPR) versus false positive rate (FPR) (Equations D.8, D.9) that illustrates the relative tradeoff between benefits (true positives, TPs) and costs (false positives, FPs). Precision-recall (PR) curves depict
the tradeoff between precision, which is a measure of exactness or quality (i.e. how many positive claims are correct), and Recall, which is a measure of completeness or quantity (i.e. how many positives were claimed of all possible positives) (Equations D.10 and D.11) [99].

\[
TP \text{ Rate} = \frac{TP}{P} \tag{D.8}
\]

\[
FP \text{ Rate} = \frac{FP}{N} \tag{D.9}
\]

\[
Precision = \frac{TP}{TP + FP} \tag{D.10}
\]

\[
Recall = \frac{TP}{P} \tag{D.11}
\]

To ensure that network inferences were robust and to assess any evidence of overfitting, we performed four-fold gold standard and genome cross validation, leave-out-one cross validation, and bootstrapping as previously described [1].

**Gold Standard Evaluation**

To assess the importance of training the Bayes net using a curated, cell-type-specific gold standard, we generated a series of test gold standards as previously described [1]. We trained Bayes nets using these alternative gold standards and a mESC data feature set, and used the performance metrics described earlier to evaluate results.
Network Topology Analysis

To analyze the network topology and evaluate biological information contained within graph files, we calculated degree ($k$), sum of degrees ($k_i$), mean degree ($k_{\text{mean}}$), and scaled degree ($K_i$), for each gene in the training set (Equations D.12–D.15) [113].

$$k_i = \sum_{j \neq i} A_{ij}$$  \hspace{1cm} (D.12)

$$k_{\text{max}} = \max(k)$$  \hspace{1cm} (D.13)

$$k_{\text{mean}} = \frac{k_i}{n}$$  \hspace{1cm} (D.14)

$$K_i = \frac{k_i}{k_{\text{max}}}$$  \hspace{1cm} (D.15)

Where the degree of the $i^{th}$ node of vector $k$ ($k_i$) equals the sum of edge weights between node $i$ and all other nodes in the training set, and Adjacency matrix $A_{ij}$ quantifies the connection strength from node $i$ to node $j$; the mean degree ($k_{\text{mean}}$) is the degree $k_i$ divided by the total number of nodes $n$ in the training set; and $k_{\text{max}}$ is the maximum degree across all $n$ components of vector $k$.

Functional Correlation Scores

For functionally directed analyses, we calculated a functional correlation score $S_i$ for a gene $i$ as the average edge weight between gene $i$ and all genes within a functional set of genes $G$ within a network represented by adjacency matrix $A_{ig}$ (Equation D.16).

$$S_i = \frac{1}{G} \sum_{g \in G} A_{ig}$$  \hspace{1cm} (D.16)
For our results, we created two sets of these scores. The first used the set of 365 genes in the positive gold standard to calculate functional correlation scores to the positive gold standard. We then used an quasi-active-learning approach to refine our set to a highly correlated subset of self-renewal genes, those with an gold standard functional correlation score of 0.25 or higher (173 mESC genes; 111 hESC genes) Using these subsets of strongly correlated genes, we calculated and scaled an updated functional correlation score to this “golden” gold standard set of known self-renewal genes. We refer to this measure as the self-renewal correlation (SRC) score. We used these SRC values to identify major gene hubs within networks and segregate clusters of genes that shared similar network properties.

**Differential Network Analysis**

We produced a simple differential network of the mESC and hESCs networks by mapping gene pairs in both networks to MGI systematic IDs (MGI:ID), sorting each network file by gene 1 and 2, then subtracting the posterior edge weight for gene pairs (mESC score minus hESC score) to evaluate genes strongly predicted to be associated in one network over the other. We performed differential correlation of mESC and hESC networks by creating adjacency matrices of functional subsets of genes associated with self-renewal and other functional processes and calculating functional correlation and dot product similarity profiles. To determine the conservation of edge weights between networks, we used the used the Spearman correlation coefficient as the Pearson correlation coefficient between ranked variables for a sample size of edges (Equation D.17). To determine whether the value of \( \rho \) was significantly different from zero, we used a Fisher transformation (Equation D.18) to calculate a \( z \)-score (Equation D.19).
\[ p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \]  

(D.17)

where \( p \) is the Pearson correlation coefficient between ranked variables for a sample size of \( n \) for the \( i \)th instance of vectors \( x \) and \( y \).

\[ F(r) = \frac{1}{2} \ln \frac{1+r}{1-r} = \arctan h(r) \]  

(D.18)

where \( F(r) \) is the Fisher transformation of \( r \), the Spearman rank correlation coefficient, and \( n \) is the sample size (in this case the total number of edges in the sample).

\[ z = \frac{n - 3}{1.06} F(r) \]  

(D.19)

Where \( z \) is a z-score for \( r \), which approximately follows a standard normal distribution under the null hypothesis of statistical independence (\( \rho = 0 \)).
BIOGRAPHY OF THE AUTHOR

Karen Dowell was born in the Philippines and graduated from high school in Concord, California. She received a B.A. in History (1980) from the University of California at Berkeley and a B.S. in Biological Sciences (2008) from the University of Southern Maine (USM). Between degrees, she spent more than 15 years working in the software industry, first as an international editor for *PC World* magazine during the early days of personal computing, and later as a marketing and technical communications expert at Data Design Associates and PeopleSoft. In her last position at PeopleSoft, she was a vice president of communication services, responsible for a team of ~60 writers and graphic designers who collectively produced more than 350 product guides and a broad range of marketing and educational materials. She left PeopleSoft to start her own publishing company and marketing communications consulting firm: Two Dog Press. In 2003, one of her Labrador retrievers died of an aggressive cancer less than a month after he was diagnosed. His death prompted Karen to return to school to learn more about cancer biology. Professors at USM ignited her interest in the biomedical sciences as well as comparative and functional genomics. And the rest, as they say, is history.

Karen received PhD training at University of Maine Graduate School of Biomedical Sciences and Engineering at The Jackson Laboratory, where she studied under the direction of Dr. Matthew Hibbs. Upon satisfactory completion of all requirements for the degree, Karen will be joining the research staff at the Department of Computer Science at Dartmouth College to work as a postdoctoral scholar in computational immunology in Dr. Christopher Bailey-Kellogg’s laboratory. Karen is a candidate for the PhD degree in Biomedical Sciences from the University of Maine in August 2013.