STEMNESS REVISITED: A META ANALYSIS OF STEM CELL SIGNATURES USING HIGH-THROUGHPUT DATA INTEGRATION

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

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

by

Martina I. Koeva

December 2009

The Dissertation of Martina I. Koeva is approved:

______________________________
Professor Joshua Stuart, Chair

______________________________
Professor Camilla Forsberg

______________________________
Professor Kevin Karplus

______________________________
Tyrus Miller
Vice Provost and Dean of Graduate Studies
Copyright © by

Martina I. Koeva

2009
# Table of Contents

List of Figures ............................................ vii
List of Tables ............................................... x
Abstract ......................................................... xii
Dedication ......................................................... xiv
Acknowledgments ............................................... xv

1 Introduction ..................................................... 1
  1.1 Motivation and problem statement ......................... 1
  1.2 Goals of the study ........................................ 3
  1.3 Overview of main results .................................. 3
  1.4 Organization of the dissertation .......................... 5

2 Stem cells ..................................................... 6
  2.1 Normal stem cells .......................................... 7
    2.1.1 Definitions, origins and key properties of normal stem cells .. 7
    2.1.2 Normal stem cell types .................................. 10
    2.1.3 Known functional pathways that regulate normal stem cell behavior 17
  2.2 Relationship between stem cells and cancer ............... 23
    2.2.1 Cancer and tumor heterogeneity ....................... 23
    2.2.2 Cancer evolution theories .............................. 25
    2.2.3 Cancer stem cells ...................................... 28
    2.2.4 Shared mechanisms between normal and cancer stem cells .......... 30
    2.2.5 Metastasis and stem cells .............................. 32
  2.3 Side populations ........................................... 35
  2.4 High-throughput technologies and stem cells ............. 36
    2.4.1 High-throughput technologies .......................... 36
    2.4.2 High-throughput data ................................... 40
    2.4.3 Public data repositories ............................... 43
9 Conclusion

9.1 Discussion .................................................. 202
9.2 Future directions ........................................... 205
  9.2.1 Application to human data .......................... 205
  9.2.2 Application to alternative splicing and miRNA data .... 205
  9.2.3 Addition of niche data ................................ 206
  9.2.4 Methodological improvements ...................... 206
  9.2.5 Possible biological experiments .................... 207

A Definitions of key terms .................................. 209

B Tables of stemness modules ............................... 213

Bibliography .................................................. 218
# List of Figures

2.1 Hierarchical organization of stem cell types within the ectoderm, mesoderm, and endoderm lineages. .................................................. 8  
2.2 Stem cells can undergo symmetric or asymmetric cell division. ......... 10  
2.3 Hematopoietic stem cell system differentiation tree. .......................... 12  
2.4 General scheme of the fluorescence-activated cell-sorting (FACS) flow cytometry technique. ................................................................. 13  
2.5 Wnt pathway activation scheme. ..................................................... 18  
2.6 Notch pathway activation scheme. ................................................... 20  
2.7 TGF beta pathway example activation scheme. ................................. 22  
2.8 Clonal evolution cancer expansion model. ........................................ 26  
2.9 Cancer stem cell expansion model. .................................................. 27  
2.10 Illustration of functional self-renewal assay. ..................................... 29  
2.11 Illustration of the role of Bmi1 in self-renewal. ................................. 31  
2.12 A FACS plot for a toy example that illustrates the isolation and definition of a side population. ............................................................... 37  

4.1 Overlap between the genes upregulated in four stem cell types in three stemness founder studies. ......................................................... 54  

5.1 Overview of stemness meta-analysis method. .................................... 62  
5.2 Published gene list (PGL) input form to Stemness Meta-Analysis Method (SMA). ................................................................. 64  
5.3 Module-level view of stemness. ...................................................... 67  
5.4 Classification of modules based on cell-diversity, specificity and gene-diversity scores. ................................................................. 69  
5.5 Global inter-experiment similarities in mouse stem cell compendium. ... 74  
5.6 Global inter-experiment similarities in human stem cell compendium. ... 77  
5.7 Protein similarity network generation approach used for homolog family definition. ................................................................. 80  
5.8 Possible pitfalls of the homolog module definition methodology. ......... 80  
5.9 Pattern classification procedure used to identify stemness modules ..... 104
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.10</td>
<td>Visual illustration of the five-fold cross-validation setup.</td>
<td>109</td>
</tr>
<tr>
<td>5.11</td>
<td>Robustness of recurrence scores across cross-validation folds.</td>
<td>113</td>
</tr>
<tr>
<td>5.12</td>
<td>Precision-recall comparison of four stemness index scores.</td>
<td>116</td>
</tr>
<tr>
<td>5.13</td>
<td>Precision comparison between stemness index scores.</td>
<td>117</td>
</tr>
<tr>
<td>5.14</td>
<td>Precision-recall comparison of twelve stemness index scores, based on different feature types and elements.</td>
<td>123</td>
</tr>
<tr>
<td>5.15</td>
<td>Precision-recall comparison of the real stemness and differentiation features to a randomly selected feature set.</td>
<td>124</td>
</tr>
<tr>
<td>5.16</td>
<td>A box-plot comparison of the stemness indices of all stem cell and differentiated cell experiments in the mouse stem cell compendium.</td>
<td>126</td>
</tr>
<tr>
<td>6.1</td>
<td>Representative recurrence score distribution of modules of size 3.</td>
<td>129</td>
</tr>
<tr>
<td>6.2</td>
<td>Selection of significant recurrently upregulated homolog families in mouse stem cell compendium.</td>
<td>130</td>
</tr>
<tr>
<td>6.3</td>
<td>Distribution of the significant recurrently upregulated modules, based on input (evolutionary or functional) source.</td>
<td>130</td>
</tr>
<tr>
<td>6.4</td>
<td>Swap control of mouse stem cell data.</td>
<td>134</td>
</tr>
<tr>
<td>6.5</td>
<td>Proliferation bias control shows low bias impact of cultured cell data.</td>
<td>135</td>
</tr>
<tr>
<td>6.6</td>
<td>Selection of a significant cell-type diversity modules in mouse SMA.</td>
<td>137</td>
</tr>
<tr>
<td>6.7</td>
<td>Global overview of classes of functional and homolog recurrent modules.</td>
<td>142</td>
</tr>
<tr>
<td>6.8</td>
<td>Stem-cell-only expression pattern of a tumor suppressor family - the p53 tumor module.</td>
<td>144</td>
</tr>
<tr>
<td>6.9</td>
<td>Stem-cell-only expression pattern of transcriptional master regulators of differentiation - the Myb gene family.</td>
<td>145</td>
</tr>
<tr>
<td>6.10</td>
<td>Stem-cell-only expression pattern of Rbp gene family.</td>
<td>146</td>
</tr>
<tr>
<td>6.11</td>
<td>Functional enrichment of stemness homolog and functional gene modules.</td>
<td>148</td>
</tr>
<tr>
<td>6.12</td>
<td>Secreted Frizzled-related protein (Sfrp) family expression in stem cells.</td>
<td>151</td>
</tr>
<tr>
<td>6.13</td>
<td>Sfrp1 model of regulation in intestinal stem cells.</td>
<td>152</td>
</tr>
<tr>
<td>6.14</td>
<td>Non-metastatic expressed (Nme) family expression in stem cells.</td>
<td>154</td>
</tr>
<tr>
<td>6.15</td>
<td>Heat shock protein 70 (Hsp70/Hspa) family expression in stem cells.</td>
<td>156</td>
</tr>
<tr>
<td>6.16</td>
<td>Inhibitor of differentiation (Id) family structural elements and functional role.</td>
<td>158</td>
</tr>
<tr>
<td>6.17</td>
<td>Global similarity comparison between homolog modules and GO gene sets/KEGG pathways)</td>
<td>160</td>
</tr>
<tr>
<td>6.18</td>
<td>Comparison of the deviation from recurrence cutoff scores between homolog and Wong et al. modules.</td>
<td>162</td>
</tr>
<tr>
<td>6.19</td>
<td>Upregulation pattern of the ATP-binding cassette, subtype B (Abcb) family of proteins in differentiated and stem cells.</td>
<td>165</td>
</tr>
<tr>
<td>7.1</td>
<td>Fraction of overlap distribution between human and mouse networks.</td>
<td>170</td>
</tr>
<tr>
<td>7.2</td>
<td>Selection of significant recurrently upregulated human homolog families.</td>
<td>171</td>
</tr>
<tr>
<td>7.3</td>
<td>Comparison of the global distributions of homolog family sizes and recurrence scores in human and mouse.</td>
<td>172</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.4</td>
<td>Correlation between the recurrence cutoffs used for every module size in common between mouse and human.</td>
<td>173</td>
</tr>
<tr>
<td>7.5</td>
<td>Swap analysis of human recurrently upregulated families.</td>
<td>177</td>
</tr>
<tr>
<td>7.6</td>
<td>Cell type recurrence cutoff selection in human SMA.</td>
<td>178</td>
</tr>
<tr>
<td>8.1</td>
<td>Stemness scores for a new mouse prostate stem cell experiment from the literature.</td>
<td>191</td>
</tr>
<tr>
<td>8.2</td>
<td>Stemness scores for new mouse stem cell and differentiated cell experiments collected from the literature.</td>
<td>193</td>
</tr>
<tr>
<td>8.3</td>
<td>Stemness scores for side and non-side populations.</td>
<td>195</td>
</tr>
<tr>
<td>8.4</td>
<td>Stemness scores for cancer stem cell populations and differentiated (non-stem-cell-like) cancer cell populations.</td>
<td>197</td>
</tr>
<tr>
<td>8.5</td>
<td>A modified hematopoietic differentiation tree with L-GMP cells.</td>
<td>198</td>
</tr>
<tr>
<td>8.6</td>
<td>Stemness scores for metastatic and non-metastatic cancer populations.</td>
<td>200</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Examples of stem cell (LT-HSC) marker signatures in mouse and human. 12

5.1 List of freshly isolated (primary) mouse stem cell profiling studies used in the mouse stem cell compendium. ................................. 70

5.2 List of cultured mouse stem cell profiling studies used in the mouse stem cell compendium. ......................................................... 71

5.3 List of cultured human cell profiling studies collected in human stem cell compendium. ......................................................... 75

5.4 List of non-cultured (primary) human stem cell profiling studies collected in human stem cell compendium. ................................. 76

5.5 Summary of the distribution of mouse homolog (evolutionary) gene modules used in the stemness meta-analysis. ................................. 81

5.6 Summary of the distribution of human homolog (evolutionary) gene modules used as input to the stemness meta-analysis method. ................. 81

5.7 Summary of the distribution of the mouse functional gene modules used in the stemness meta-analysis. ......................................................... 83

5.8 Summary of the distribution of human functional gene modules used as input to the stemness meta-analysis method. ................................. 83

5.9 Overview of the recurrence-score-associated notation. ......................................................... 86

5.10 AUC results for all 13 tested recurrence scoring methods, based on synthetic data evaluation. ......................................................... 94

5.11 Recurrence score cutoffs for modules of different sizes. ................................. 97

5.12 Overview of the diversity-score-associated notation. ......................................................... 99

5.13 Overview of the stemness index score notation. ......................................................... 110

5.14 Number of stemness families identified in each cross-validation (CV) fold. 111

5.15 Number of differentiation families identified in each cross-validation (CV) fold. 112

6.1 The 38 stemness genes and significantly enriched functional categories. 139

6.2 Summary of the classification of recurrently upregulated homolog and functional gene modules. ......................................................... 141
6.3 List of all mouse differentiation homolog modules. .......................... 164

7.1 Summary of the classification of recurrently upregulated human homolog 
and functional gene modules. .................................................. 180
7.2 List of all human stemness evolutionary and functional gene modules. . 181
7.3 List of all human differentiated evolutionary and functional gene modules. 182

8.1 Distribution of the sources of mouse test populations used to predict 
self-renewal capacity. .......................................................... 189

B.1 List of all mouse stemness homolog AFA modules. ....................... 215
B.2 List of all mouse stemness functional AFA modules. ....................... 216
B.3 List of all mouse stemness homolog OFA modules. ....................... 217
B.4 List of all mouse stemness functional OFA modules. ....................... 217
Abstract

Stemness Revisited: A Meta Analysis of Stem Cell Signatures Using High-Throughput Data Integration

by

Martina I. Koeva

Stem cells are functionally defined cells with a high therapeutic potential for many diseases. The stemness hypothesis states that stem cells share a core set of mechanisms that regulate the shared stem cell properties of self-renewal and multi-lineage potential. Previous attempts to identify genes required for core stem cell function across stem cell types using transcriptional profiling have identified few such genes. My work focused on the development of a computational stemness meta-analysis (SMA) method that uses high-throughput differential gene expression data integration to address three main questions: do functional redundancy and tissue-specific expression mask common molecular mechanisms shared between stem cell types? Are stemness mechanisms conserved between mouse and human stem cells? Can we use gene expression signatures to predict stem cell state?

The SMA method identified 103 mouse evolutionarily related groups of homologous genes with reproducible, statistically significant, cell type diverse and stem cell-specific upregulation in multiple stem cell types. The results point to specific examples of functional redundancy in modules controlling cell adhesion, quiescence, and gene silencing. Shared homolog modules also include genes in the Myc, Myb, Chd, Hspa,
Id, and many other families. Genes within the stemness homolog families are prime
candidate regulators of conserved stemness mechanisms and may play critical roles as
stem cell markers.

I directly measured the level of conservation of stemness mechanisms between
mouse and human cells. Application of the SMA method to a human stem cell com-
pendium indicates that human data are globally more heterogeneous than murine stem
cell data. However, human stemness families incorporate several conserved mammalian
stemness modules, such as the Integrin α, TCF/LEF, Frizzled, Notch, and Chd families.

Finally, I used the stemness modules identified in the mouse SMA to define a
stemness index score and evaluate how stem cell-like a new gene expression signature
is. I validated the predictiveness of the stemness modules through an internal cross-
validation test and applied the stemness index test to a large set of new experiments
from normal stem cells, side populations, cancer stem cells, and metastatic populations.
The results indicate that mouse stemness modules could predict stem cell-like features
in various data sources with high accuracy.
To my family,

Stefka, Iordan and Petya,

whose brilliance, dedication and support have been an amazing source of energy and inspiration.
Acknowledgments

My stemness research was supported by a pre-doctoral fellowship from the California Institute for Regenerative Medicine (CIRM).

I would like to thank Joshua Stuart and Camilla Forsberg for their invaluable ideas, guidance, support and energy throughout this work. I have had the wonderful opportunity to work closely with them to and draw from their complementary expertise, which has been instrumental to the completion of this project.

I would also like to thank Kevin Karplus and Raquel Prado for a great learning experience, valuable feedback, as well as their patience with reading of my work.

I would like to thank David Bernick, Charles Vaske, and Alex Williams who have discussed various aspects of my research on many occasions during the last few years. Many other graduate students have provided feedback and recommendations on various aspects of the project and its presentation, including Matt Weirauch, Grant Thiltgen, Courtney Onodera, Daniel Carlin, Josue Samayoa, Daniel Sam, Thomas Juettemann, Marcos Woerhmann, Firas Khatib and others.

Thank you to Katy Elliot who has helped me on many occasions throughout my CIRM appointment and Carol Mullane for additional support and help.

Finally, thank you to my family without whose support and belief in me, this would have been harder to accomplish.
Chapter 1

Introduction

1.1 Motivation and problem statement

My research focuses on elucidating global mechanisms of stem cell function. Stem cells are functionally defined cells, characterized by their ability to self-renew and give rise to many different mature cell types. In recent years stem cells have gained much scientific prominence, because their functional characteristics make them highly relevant for therapeutic purposes in many neuro-, muscle- and other degenerative diseases. Understanding the functional and regulatory mechanisms of these cells has become even more important with the discovery of cancer stem cells, which seem to play a central role in cancer through their ability to regenerate the fully differentiated cancer cell populations generally targeted by current cancer treatments.

Normal stem cells have been identified in many major organs, tissues and systems (e.g. blood, liver, kidney, lung, muscle, etc.), while cancer stem cells have
been discovered for cancer cell types varying from leukemia to neuroblastoma. Despite their importance, both normal adult and cancer stem cells have been very hard to study because of their rare nature in humans and many model organisms. Isolation and purification of these cells for many tissue types has been a major hurdle, because of the lack of good marker genes that would allow the separation of stem cells from progenitor and fully differentiated cells in the tissue of interest.

Much work has been done to elucidate the mechanisms of stem cell function. Especially with the advent of high-throughput technologies, many large-scale data sets have been published to examine the gene expression patterns of various stem cell types in the hopes of discovering stemness genes. However, researchers have still not found a single set of individual stemness genes common to all stem cell types that allow all of these cells to self-renew and maintain their stem-like cell state.

So how do stem cells achieve their function? We know that genes do not act on their own within the cell. They often need to interact with other genes within pathways or protein complexes to achieve their designated cellular function. In relation to this observation, stem cell researchers have started to look to more global mechanisms of stem cell state regulation. Because of the stochasticity of gene expression, it is possible that even though different stem cell types do not directly use the same genes, they may share the same, or redundant functional pathways. The genome also has a built-in robustness provided by gene duplications, which suggests that evolutionarily-related genes may also provide a way for stem cells to use different genes, and yet perform the same function. If we account for these possibilities, would common stem cell mechanisms
1.2 Goals of the study

In the context of the problem stated in the previous section, my work in this dissertation aims to computationally address three important questions:

1. Do functional redundancy and tissue-specific expression mask the common stem cell mechanisms?

2. If common stem cell mechanisms exist, are they conserved between mouse and human stem cells?

3. Can we predict the state of differentiation of a cell based on its gene expression signature?

1.3 Overview of main results

To address the role of functional redundancy in stem cells, I first develop the methodology to test for global reproducible expression of entire gene sets across multiple conditions. I develop the Stemness Meta-Analysis (SMA) method that uses techniques derived from standard meta-analysis theory to identify gene sets (modules), which are significantly recurrently upregulated across many stem cell experiments, are represented in most stem cell types, and are specific to stem cells as opposed to differentiated cells. Using this method, I identify approximately 103 stemness modules of evolutionarily re-
lated homologous genes with reproducible, statistically significant and stem cell-specific upregulation in many mouse stem cell types. Genes within these homolog families are prime candidate regulators of conserved stemness mechanisms and may play critical roles as stem cell markers. They include many known self-renewal genes, such as Myc, Myb, Chd1.

To address the conservation of stem cell mechanisms between mouse and human cells, I apply the SMA method to a large compendium of stem cell data in human. Results suggest that even though much data is available for human stem cells, the gene expression signatures associated with them are much more heterogeneous than their mouse counterparts, presumably related to the lack of good marker genes for the isolation of pure populations in human. Data are also available for fewer stem cell types. Nevertheless, I find conservation of five major stemness families: the Notch, Frizzled, Chd, TCF/LEF, and Integrin α families. I expect that some of the discovered mouse stemness families can be used as putative markers in human stem cell populations.

Finally, to address the predictability of stem cells based on their expression signatures, I define a stemness index score that uses stemness modules to measure how stem cell-like a new gene expression signature is. I apply this index scoring to mouse expression signatures derived from new stem cell experiments, side populations (putatively stem-cell-enriched populations of cells with highly active transporter proteins), cancer stem cells, and metastatic populations. The results suggest that at least in mouse the stemness modules, as used by the stemness index score, are highly predictive of normal and cancer stem cell populations, as well as side and metastatic populations.
1.4 Organization of the dissertation

The dissertation is organized in nine main chapters with a separate appendix section that presents the definitions of key italicized terms used in the main text. The current chapter (Chapter 1) provides some brief motivation for the study along with some basic definitions. It also describes the central problem addressed by the current work and gives a summary of the significance of the results. The purpose of Chapter 2 is to provide some of the stem cell- and cancer-related biological background needed to understand the main results, discussed in later chapters. Chapter 3 is focused on the introduction of meta-analysis, as well as its relevance and application to high-throughput data integration. This chapter is integral to the understanding of the methodology used in this study. Chapter 4 introduces the work directly relevant to the study of common stem cell mechanisms, including previous attempts at a molecular definition of stemness. A detailed presentation of all methods designed for this study is presented in Chapter 5. The subsequent three chapters focus on results and applications: Chapter 6 discusses stemness mechanisms in mouse, Chapter 7 identifies stemness mechanisms in human and discusses the similarities and differences between mouse and human, while Chapter 8 shows some interesting applications of the discovered stemness mechanisms to the study of metastasis, cancer stem cells and stem cell niches. The final chapter (Chapter 9) provides a discussion of some of the major implications of this work and describes directions for future work. Definitions of key terms are included in Appendix A.
Chapter 2

Stem cells

This chapter aims to provide the biological background needed by the reader to understand the basic mechanisms of stem cell function. Section 2.1 illustrates some basic stem cell definitions and key properties of normal stem cells. It also introduces important stem cell concepts, describes the most common stem cell types and examines some known pathways associated with self-renewal. Section 2.2 introduces different cancer evolution theories and discusses the relationship between normal stem cells, cancer stem cells and metastasis. Section 2.3 describes side populations and how they relate to stem cells, and finally Section 2.4 introduces various high throughput technologies used for measurement of gene expression and discusses their use in stem cell analysis.

An understanding of the stem cell background provides the context for the interpretation of the results of my work, and gives the reader ideas for potential positive controls. While the input data used in my study of stemness vary substantially, the underlying theme behind all input types is their tie to stem cells and stem-cell-like
2.1 Normal stem cells

2.1.1 Definitions, origins and key properties of normal stem cells

Stem cells are undifferentiated cells characteristic of multi-cellular species. They have two characteristic properties that functionally define them as stem cells: the ability to differentiate to many different mature cell types and the ability to self-renew, or produce more stem cells of their own type. [192]. At the broadest level of classification, stem cells can be characterized as either embryonic stem cells (ESC) or adult stem cells (ASC).

Early in embryogenesis before implantation, the blastocyst develops two separate components: the inner cell mass (ICM) and the trophoderm [128]. The trophoderm forms the outside of the embryo and its trophoblast cells are the ones that are destined to give rise to all extra-embryonic tissues and form the placenta. The inner cell mass consists of the cells that will eventually develop into all cells and tissues of the embryo. It is from the ICM that embryonic stem cells are derived. Thus, ES cells are pluripotent in nature and can give rise to any of the three developmental germ layers – ectoderm, mesoderm, or endoderm – and develop into any cell type in the organism with the exception of extra-embryonic tissues. This quality makes them uniquely valuable and highly potent for therapeutic purposes.

The extent of self-renewal of embryonic stem cells in vivo is not very well-
understood, but *in vitro* both human and mouse ES cells have the ability to extensively proliferate and self-renew for an unlimited amount of time [115, 161]

Unlike embryonic stem cells, adult stem cells (ASC) have to maintain cells and tissues for the life span of an organism. Even though they can self-renew and give rise to more stem cells of the same type, they are generally *multipotent* in nature and can only generate more mature cells of the same system or tissue type. Specifically, within the ectoderm lineage neural stem cells will give rise to all the cells of the nervous system, while epithelial (skin) stem cells will develop and maintain the skin of the animal (Figure 2.1). Similarly, within the mesoderm lineage hematopoietic stem cells will support the development of the bone marrow, the blood system and all its myeloid and lymphoid components, while mesenchymal stem cells will be involved in the development of the support cell system that contributes to stem cell maintenance. Other mesoderm-derived
tissue-specific stem cells will give rise to the muscle and bone tissues. Finally, within the endoderm lineage there are organ-specific stem cells, such as lung, liver, gastric and intestinal stem cells that can give rise to the different cells that make up each of these corresponding organs (respectively lung, liver, stomach and intestines).

Based on their properties, stem cells can generally fall into one of several states: *quiescence, proliferation, differentiation* or *apoptosis*. Unlike most normal cells, stem cells have the ability to enter into a special state of the cell cycle, called $G_0$, and only maintain their stem cell state with the support of the appropriate stem cell niche. This state is generally referred to as the *quiescent* state. The more active states of proliferation and differentiation are most relevant in wound repair or for the general maintenance of tissues with a fast cell turnover.

*Proliferation* and self-renewal require that stem cells enter an active state. Dependent on the needs of the organism or the tissue type, stem cells can undergo either symmetric or asymmetric divisions (Figure 2.2). In asymmetric divisions, a stem cell gives rise to two different daughter cells: a stem cell and a more differentiated cell. This process ensures that stem cells are not lost, while differentiated cells are generated. In symmetric divisions, a stem cell gives rise to either two stem cells, or two more differentiated cells.

*Differentiation* is another active state important for a stem cell and the breadth of differentiation potency is one of the major factors that determines how a stem cell is defined. Besides pluripotent and multipotent stem cells, there are also cells that have a limited ability to differentiate into only one or several mature cell types and have
Stem cell

Differentiated cell

Symmetric division

Asymmetric division

Figure 2.2: Stem cells can undergo symmetric or asymmetric cell division. In symmetric division, the stem cell (red) can produce either two stem cells, or two more differentiated (green) cells. In asymmetric division, the stem cell can produce one stem cell and one more differentiated cell.

a very limited self-renewal ability. These cells are referred to as progenitor cells and even though their functional properties partially overlap with stem cells, they are not generally considered such.

Finally, stem cells can also undergo apoptosis under certain conditions in a fashion similar to other normal cell types.

2.1.2 Normal stem cell types

2.1.2.1 Hematopoietic stem cells

The hematopoietic system is one of the best-studied systems of stem cells in both human and mouse. Its structure can be most easily represented and visualized as a hierarchically organized tree, where the stem cell is at the root and the leaves
represent the most committed cells with no self-renewal capacity (Figure 2.3). At the
top of the cellular and differentiation hierarchy is the long-term hematopoietic stem cell
(LT-HSC), which can give rise to a short-term hematopoietic stem cell (ST-HSC) and
can also proliferate generally for the life span of the organism [116]. ST-HSC cells can
give rise to a multipotent progenitor cell (MPP) and retain some renewal capabilities,
while MPP cells only have the ability to differentiate into any myeloid or lymphoid
cell type without a considerable self-renewal potential. They can give rise to two more
lineage-committed types of progenitor cells: common lymphoid progenitors (CLP) [98]
and common myeloid progenitors (CMP) [2], which as their names indicate can give rise
to all types of blood cells within their lineage.

To study the functional features of these different cell populations independ-
dently, scientists generally use cell surface marker genes that collectively can distinguish
between individual cell types. Cell surface markers are particularly well-understood in
the hematopoietic system and are used along with other features to discriminate between
different types of hematopoietic stem, progenitor and lineage-specific (myeloid/lymphoid)
cells. Such markers include CD13, CD19, CD34, CD38, CD45 (B220), CD71, CD133,
c-kit, Mac-1, etc.

We can describe or summarize the state of a cell by the expression of its
markers and examples of marker signatures are shown in Table 2.1. The marker genes
can then be used for purification purposes to sort different populations through an
experimental method, called fluorescence-activated cell-sorting (FACS) flow cytometry
(Figure 2.4). The resulting products are well-separated cell populations, which can then
Figure 2.3: The hematopoietic stem cell system differentiation tree shows the entire differentiation hierarchy of cells in the blood system. The most multipotent cell – the LT-HSC cell – is at the top of the hierarchy and can give rise to all cells downstream of it. The leaves of the tree represent the most committed mature cells with no self-renewal capacity.

<table>
<thead>
<tr>
<th>Organism and cell type</th>
<th>Marker signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse LT-HSC</td>
<td>CD34&lt;sup&gt;−&lt;/sup&gt; Sca-1&lt;sup&gt;+&lt;/sup&gt; Thy1.1&lt;sup&gt;+/−&lt;/sup&gt; CD38&lt;sup&gt;+&lt;/sup&gt; c-kit&lt;sup&gt;+&lt;/sup&gt; lin&lt;sup&gt;−&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>CD48&lt;sup&gt;−&lt;/sup&gt; CD150&lt;sup&gt;+&lt;/sup&gt; CD248&lt;sup&gt;−&lt;/sup&gt;</td>
</tr>
<tr>
<td>Human LT-HSC</td>
<td>CD34&lt;sup&gt;+&lt;/sup&gt; CD38&lt;sup&gt;+/−&lt;/sup&gt; Thy1.1&lt;sup&gt;+&lt;/sup&gt; c-kit&lt;sup&gt;+&lt;/sup&gt; lin&lt;sup&gt;−&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Table 2.1: Examples of LT-HSC marker signatures in mouse and human used in fluorescence-activated cell-sorting (FACS) to sort cells. The first column shows the organism and cell type for each signature, while the second column contains the signature used for sorting of the cells. The “+” superscript next to a gene indicates a cell of that type is positive for that marker, “−” indicates a cell is negative for that marker, and “lo” indicates a cell expresses low levels of that marker.
Figure 2.4: General scheme of fluorescence-activated cell-sorting (FACS) flow cytometry technique: cells are bound to fluorescently tagged antibodies and are measured one by one in a stream by the flow cytometer. When each cell is scanned, the dye associated with its bound antibody emanates color, which is captured by the instrument. Each cell is then positively or negatively electrically charged based on the color and sorted into a separate population.
be used in comparative or other experiments. Knowledge of cell surface markers has been tremendously useful in the field of stem cell biology and the study of the functional properties of individual stem cell types. However, many systems lack an abundance of well-known markers, which shows the urgent need for the discovery of new marker genes in these systems and tissue types.

2.1.2.2 Neural stem cells

Neural stem cells, like hematopoietic stem cells, have also been widely studied [57], but their mechanisms are much less understood than those of hematopoietic cells, partially because of the different nature of the nervous system from the blood system. The mammalian brain consists of a large number of structures and it is not clear whether stem cells isolated from the independent components may have a different differentiation potential. This observation is also relevant for differences between the adult and embryonic brain [57].

Neural stem cells can be isolated from the subventricular zone (SVZ) of the adult mammalian brain or alternatively from various other locations in the central and peripheral nervous system of the developing mammalian brain. Most neural stem cells are studied in vitro and can be cultured in the presence of either EGF or FGF2 and subsequently differentiated into various mature neural cell types [57], which include various types of neurons, astrocytes, oligodendrocytes and others. Some of these lineages are well-understood and appropriate markers exist for each cell or lineage type, while others still remain elusive. Specifically, neural stem cell self-renewal is dependent on the
expression of Bmi-1 [113] and many neural progenitor cells express Nestin. Similarly, Tuj1 is a general neuronal marker [59], GFAP (glial fibrillary acid protein) is an astrocyte marker [191], and GalC is an oligodendrocyte marker [21].

2.1.2.3 Embryonic stem cells

Embryonic stem cells are perhaps the most extensively studied type of stem cells, because of their readiness for culture and expansion *in vitro*. One set of commonly expressed transcription factors and marker genes truly characteristic of embryonic stem cells that regulate some of the initial self-renewal and cell fate choices of ES cells include Nanog, Oct-4 and Sox2. The role of these transcriptional factors is to suppress the expression of genes required for the process of differentiation[23].

As previously discussed, embryonic stem cells are pluripotent in nature and can give rise to any cell or tissue type in an organism. Most recently, a different set of *induced pluripotent stem (iPS) cells* that display the hallmark features of embryonic stem cells has also been identified in both human and mouse [171, 170, 130]. The reprogramming or induction of pluripotency features usually begins with a primary fibroblast cell and the integration and activation of several transcription factors. The initial set consisted of four different transcription factors including two oncogenes (Oct4, Sox2, Klf4 and Myc), but since then various other combinations and subsets of factors have been found to also induce the reprogramming of the fibroblast cells into ES-like pluripotent cells.
2.1.2.4 Intestinal stem cells

Intestinal stem cells (ISC) are adult stem cells that give rise to the various lineages of cells that make up the lining of the intestinal epithelium. There are four different cell types that can be generated by intestinal stem cells: Paneth cells, goblet cells, enteroendocrine and columnar enterocyte cells [16, 72, 107]. The intestinal stem cells can be found above the Paneth cells, which are situated at the base of the crypt – the location of the intestinal stem cell niche. They express a variety of genes, a few of which can be used as markers, such as Noggin, Tcf4, Ephb3, and Musashi-1 [13, 18, 69, 99, 120]. However, most of them are not exclusive to intestinal stem cells. Other markers include the hematopoietic stem cell marker Bmi-1, which has most recently been identified to be expressed in intestinal stem cells as well [150].

2.1.2.5 Epithelial stem cells

Epithelial stem cells can be found in the hair bulge of animals, underneath the sebaceous gland and reside in their stem cell niche in a quiescent state. After activation, the stem cells can give rise to either stem cells and progenitors that continue to reside in the bulge, or to progenitor cells that transition upwards closer to the skin surface, where they can be used to generate skin (epidermal) cells for wound repair or general epidermal maintenance. Alternatively, the daughter progenitor cells can transition down to the hair matrix, where they can differentiate and develop into hair shaft cells [107, 127, 174].
2.1.3 Known functional pathways that regulate normal stem cell behavior

Given the variety of stem cell types, one naturally wonders: if all of these cells need to perform similar roles in their environment, could they have any mechanisms in common? What mechanisms do stem cells use to self-renew and how do they maintain the fine homeostatic balance between quiescence, proliferation, differentiation and apoptosis?

Several functional pathways have emerged over the years in the context of their role in various stem cell types. I next describe some of the most important functional networks known to regulate stem cell behavior in at least a few stem cell types. In the study of stemness, these networks – Wnt, Notch and TGFβ – are especially important as they are the most likely stemness candidates.

2.1.3.1 Wnt pathway

The Wnt pathway is activated when a Wnt ligand, one of the many molecules within the Wnt family of proteins, binds its partner receptor – Frizzled – at the cell surface (Figure 2.5). This binding event activates a signal across the membrane and the intracellular Disheveled gene becomes itself activated, which negatively regulates the machinery responsible for the degradation of β-catenin in the cell. The degradation complex consists of several molecules including APC, GSK3β and Axin/Compactin. In the absence of this degradation complex, β-catenin is free to accumulate and translocate to the nucleus, where it can interact with the TCF/LEF transcription factor family.
Figure 2.5: Wnt pathway activation scheme. The Wnt ligand binds a Frizzled (Fzd) receptor on the cell surface, which activates the Disheveled (Dsh) gene. Dsh inhibits the activity of the degradation complex, which is responsible for the degradation of β-catenin. In the absence of the degradation, β-catenin accumulates and activates TCF/LEF in the nucleus. TCF/LEF factors regulate the activity of self-renewal genes.

members, which in turn activate many proliferation, self-renewal, and other stem-cell related genes. The mode of action of the Wnt pathway is similar in most stem cell types and the activation of this network is generally associated with self-renewal[183].

Wnt plays a significant role in embryonic stem cell self-renewal, as observed in many biological experiments. Activation of the pathway can experimentally occur either through the overexpression and binding of a Wnt ligand (e.g. Wnt1) to the Frizzled receptor, allowing the accumulation of β-catenin in the cell, or alternatively through the direct inhibition of the components of the degradation machinery also resulting in the accumulation of β-catenin [183]. Both of these techniques have been directly used and results indicate that the accumulation of β-catenin generally results in the inhibition of differentiation [8, 68] and the persistence of the undifferentiated state[151].
This pathway also has a central role in the regulation of stem cell expansion and differentiation in hematopoietic and epithelial stem cells. In cultured HSCs the addition of Wnt ligand, as well as the direct overexpression of \( \beta \)-catenin result in HSC expansion and prolonged HSC self-renewal. Other work suggests that overexpression of one of the degradation complex components, Axin, has a negative effect on the growth of HSCs [9, 144]. Similarly, in culture conditions \( \beta \)-catenin has a significant effect on the number of epithelial stem cells and the pathway has been shown to directly modulate epithelial stem cell fates and regulate self-renewal and differentiation in the skin [111, 201].

2.1.3.2 Notch pathway

Another pathway with an important role in the regulation of stem cell self-renewal and differentiation is the Notch pathway (Figure 2.6). The central molecules of this pathway are the members of the Notch family of receptors — Notch1–4. These receptors can be activated by several groups of ligands that include the Delta-like family of proteins, as well as the family of Jagged proteins. Once binding of the ligand to the extracellular component of the Notch receptor occurs at the cell surface, the Notch intracellular domain cleaves and moves to the cell nucleus. There it binds CSL (or Lag1), which is associated with the transcriptional activity of the target genes of Notch. When Notch is inactive, CSL leads to the transcriptional inhibition of Notch targets, but in the presence of the Notch intracellular domain, CSL associates with activators instead of repressors and the Notch target genes (including Hes1 and Hey1) are activated. Many of these target genes are themselves repressors of different lineage-specifying genes and
Figure 2.6: Notch pathway activation scheme. A Delta-like (Dil) or Jagged (Jag) protein binds the Notch receptor and upon activation, the intracellular domain of the Notch receptor (NID) cleaves and relocates to the nucleus. There, NID binds to CSL and functions to activate many Notch targets, which often function as repressors of lineage-specifying genes.
Notch at a global level plays a role in self-renewal [183].

For example, expression of Notch1 has been shown to increase the level of HSC self-renewal in vitro, reduce the differentiation capacity of hematopoietic cells, as well as guide more committed progenitors towards T-cell development, rather than the B-cell lineage [183]. The Notch pathway also regulates stem cell self-renewal in the neural system: Notch activation promotes self-renewal and expansion of neural stem and progenitor cells, as well as inhibition of neural differentiation into particular lineages (oligodendrocytes, glial and others)[118, 124, 188]. Inhibition of the downstream targets of Notch, such as Hes1 has been shown to lead to depletion of the stem and progenitor pool and acceleration of differentiation [78].

2.1.3.3 TGFβ pathway

The transforming growth factor (TGF)β signaling pathway plays a central role in the regulation of proliferation and differentiation in embryonic and adult stem cells. The TGFβ superfamily of proteins consists of three major components: the TGFβ proteins, the activins and the largest set, the bone morphogenic proteins (BMP). Each of these molecules has the ability to signal into the cell through binding to a cell surface receptor, as follows: the signaling molecules bind to a type II TGFβ receptor (Figure 2.7) [5].

This interaction activates the receptor and it associates with the other receptor component – the type I TGFβ receptor. This process promotes the phosphorylation and activation of any of the five possible regulatory receptor-activated SMAD ("moth-
Figure 2.7: TGFβ pathway example activation scheme. When the TGFβ ligand binds to the type II receptor (1), the type I receptor is activated, the inhibitory I-Smads – Smad 6/7 are displaced (2), the R-Smad Smad2 is recruited (3) and activated through phosphorylation (4). Smad2 binds to the Co-Smad protein, Smad4 (5) and the complex moves into the nucleus (6), where it can interact with other proteins to activate the TGFβ targets (7).

Smad2 and Smad3 are activated by TGFβ and activins, while Smad1, Smad5, and Smad8 are phosphorylated upon BMP activation. Once phosphorylated, they bind the co-activator SMAD protein, Smad4, and move into the nucleus, where they can associate with other molecules and bind DNA to activate BMP or TGFβ target genes. Two other SMAD proteins exist – the inhibitory Smad6 and Smad7 – and are known to interact with the type I receptors to regulate the activation and de-activation of this pathway [183].

In mouse embryonic stem cells, BMP signals actively prevent neural differentiation and maintain ESC self-renewal [189]. After application of SB-421542, an inhibitor of the TGFβ signaling molecules [77], mouse embryonic stem cells showed decreased levels of ESC markers in human [82, 184], as well as lower proliferation levels in mouse
[123]. In some somatic stem cells, such as intestinal and epithelial (hair) stem cells, activation of TGFβ signaling is associated with stem cell quiescence, as BMP is thought to counteract and balance the Wnt signaling pathway by limiting and reducing the amount of β-catenin available in the cell [189]. In other stem cell types, such as mesenchymal stem cells, however, TGFβ can also positively regulate proliferation [189].

2.2 Relationship between stem cells and cancer

One of the major roles of normal stem cells in an organism is to replenish the supply of cells constantly needed to maintain all tissues, organs, and systems in “a good running condition” as well as respond in emergency conditions. In normal homeostatic conditions the cells within a tissue or an organ interact and maintain a balance between growth, differentiation, and cell death. Balance is achieved through the activation and interaction of a system of regulatory molecules and a well-developed cell communication network. Thus, a disruption of this well-organized network of molecular and cellular interactions can have serious consequences and lead to various diseases among which one of the most serious, well-studied and yet poorly understood diseases is cancer [190].

2.2.1 Cancer and tumor heterogeneity

Cancer is a disease of unchecked cellular growth and is currently among the leading causes of death in the United States. Each individual growth of abnormal cells is called a tumor and arises from normal cells and tissues. Tumors are monoclonal in nature [190], which means they arise from a single common abnormal “ancestral” cell,
as opposed to multiple different cells that could have independently undergone cancer-related mutations to give rise to the highly heterogenous assembly of cells that a tumor often contains.

Tumors can be divided into primary and metastatic, based on the location of the original abnormal changes. The former are tumors that have developed at the site of origin of the ancestral cell mutation, while the latter are tumors that have evolved from a primary tumor, but have subsequently invaded a more distant and often unrelated to the primary tumor location. The frequently fatal consequences of metastasis have led researchers to study the mechanisms, which guide the development of metastatic cancer [190].

Within the general cancer categories, the distinct cancer types and subtypes can present much heterogeneity. To date, more than 80% of tumors that afflict the population are epithelial-derived cancers, known as carcinomas and these include cancers of most internal organs, such as the stomach, small intestine, large intestine, liver, lung, pancreas. They also include ovarian, testicular, bladder, breast and skin cancers, as well some other more rare types.

Understanding the sources of heterogeneity is key to the development of new treatments and methods. Tumor heterogeneity has been observed at several different levels, but the heterogeneity that is most relevant to stem cell analysis is the one observed at the cellular level within a single tumor growth [71]. Specifically, each individual tumor can be a highly phenotypically heterogeneous assembly of cells. The cells that make up the tumor can differ in their morphology, differentiation status, proliferation potential,
the cell surface markers they express, their pathology, as well as other factors [71]. From the monoclonal tumor origin perspective this type of heterogeneity is most interesting.

Why is it that tumors that arise from a single cell are so heterogeneous in nature? A potential answer to this question lies within the stem cell field, where in at least some organs and tissues a single stem cell can be sufficient to regenerate the entire system of diverse and narrowly-specialized differentiated cells. This link between normal stem cells and cancer can give us clues as to how cancer cells can evolve to such a heterogeneous state.

### 2.2.2 Cancer evolution theories

Two main cancer evolution theories have been developed through the years: the *clonal expansion* and *cancer stem cell* models.

In the clonal expansion model all cells are equal and each one has the potential to give rise to a new tumor. Tumor development begins with the mutation of a random cell that gives it the advantageous ability to proliferate fast and dominate the surrounding tissue (Figure 2.8). Because of the unchecked growth, the cell and its progeny soon dominate the cellular neighborhood sufficiently and accumulate another mutation. The new mutation allows these cells to grow and expand even more, which inadvertently makes them even more susceptible to becoming targets of the next mutation. Each mutation hit and subsequent cellular growth can be described as an *expansion*. Within several such clonal expansions, the cells have grown and mutated sufficiently to be considered the origins of development of a new tumor [186, 190].
Clonal evolutional model

Figure 2.8: Clonal evolution cancer expansion model. The clonal expansion model assumes that all cells within a tumor have equal proliferative ability. Once an initial mutation (MUT1) occurs within a cell (dark red), the cell gains proliferative potential and its progeny quickly dominates the surrounding tissue. This event predisposes the already mutated cells to a second hit (MUT2; pink cell) and after the subsequent expansion to a third hit (MUT3; blue cell). After several rounds of expansion and accumulation of new mutations, a tumor can arise from the mutated cells.
Figure 2.9: Cancer stem cell expansion model. The cancer stem cell model assumes that some cells within a tumor have a higher proliferative and self-renewal ability than other cells. If a mutation (MUT1) occurs in a normal stem cell that causes it to become uncontrollably proliferative, the new cancer stem cell (dark green) can give rise to many cancer progenitor (green) and more differentiated cancer cells (yellow). Alternatively, if a mutation occurs in more differentiated cell, the event will only cause an extensive expansion if the mutation (MUT2) caused the cell to de-differentiate to a more stem cell-like state.

The cancer stem cell model postulates that all cells are not created equal and some cells – the cancer stem cells (CSCs) – have a higher proliferative potential than others. Similarly to the normal stem cell hierarchy, tumor cells also have a hierarchical organization, where only the cancer stem cells can give rise to new cancer cells and allow tumor proliferation\(^1\) (Figure 2.9). Most differentiated cancer cells have a limited proliferative potential and capacity and are incapable of promoting tumor growth [186].

\(^1\)It should be noted that the CSC is not necessarily the cell that was originally hit by the mutation that prompted the beginning of the oncogenic process.
2.2.3 Cancer stem cells

The existence of cancer stem cells was first shown by John Dick in an acute myeloid leukemia (AML) study, which aimed to compare the similarities between normal hematopoietic stem cells and leukemic cells. The researchers in this study used a self-renewal assay technique [37] that has been widely used as the method for discovery of new cancer stem cells. They used the markers associated with normal human HSCs (CD34+ CD38neg Lin−) to sort and distinguish between two leukemic subpopulations - a minority subpopulation of CD34+ CD38neg and a majority population of CD34+ CD38pos cells. Cells from both populations were then independently injected into severely immunocompromised mice and the level of engraftment and reconstitution of the cancer were subsequently measured. Only the cells derived from the minority subpopulation had the ability to proliferate and give rise to the heterogenous assembly of cancer cells that resembled the original cancer population [41]. This study showed the existence of a small number of cells – the CSCs – with an enormous proliferative capacity, sufficient to reconstitute a cancer population.

The self-renewal assay (Figure 2.10) used for discovery of new cancer stem cells has not changed substantially and requires four main steps: selection of cell surface markers that allow cellular subsampling, separation of the cells of interest into subsamples using FACS technology based on marker expression, independent injection of cells from each subsample into immunocompromised mice, and measurement of self-renewal capacity in the host. This technique has now been used to discover cancer stem
Fluorescently label cells

Laser beam

CSC self-renewal and expansion capacity

Figure 2.10: Illustration of functional self-renewal assay. The first step of the assay requires the selection cell surface markers and the fluorescent labeling of cells. FACS flow cytometry is then used to sort the cells based on the expression state of the markers. Each purified sub-sample can then be injected into severely immuno-compromised mice. If the sub-sample contains primarily putative cancer stem cells (red cells), a tumor growth will develop because of the high self-renewal and expansion capacity of these cells. Alternatively, if the sample does not contain cancer stem cells, but more differentiated cancer cells (green and/or white), no tumor growth will develop because of the low proliferative capacity of these cells.
cells in both breast [4, 129] and brain [158] tissue, as well as other cell types.

Al-Hajj et al. [4] discovered a small subpopulation of CD44+CD24− cells with an extraordinary proliferative capacity when injected into SCID mice in samples from breast cancer patients. Specifically, approximately 100-200 of these cells were sufficient to induce tumor reconstitution, while several fold higher quantities of other cells were insufficient and incapable of achieving the same. The CD44+CD24− cells had both the ability to self-renew and to give rise to differentiated cancer cells, exhibiting the functional properties of a stem cell [4].

Similarly, Singh et al. [158] observed that in glioblastomas a small population of cells defined by CD133+ expression also had very high proliferative capacity, unlike the rest of the tumor and could give rise not only to more cells of the same phenotype, but also to other cell types in a manner similar to the one observed in the original tumor [129, 158].

### 2.2.4 Shared mechanisms between normal and cancer stem cells

As outlined in the earlier sections, evidence has accumulated to suggest that normal and cancer stem cells share some functional properties, namely the abilities to self-renew and differentiate. Shared function indicates that normal and cancer stem cells may also share functional pathways that regulate self-renewal and differentiation and some evidence already exists to support this hypothesis.

One of the best examples of molecules and pathways that regulate both normal and cancer stem cells is Bmi-1. Bmi-1 is a marker of self-renewal in both normal
Figure 2.11: Illustration of the role of Bmi1 in self-renewal. Bmi-1, also known as Pcgf4, functions to suppress the cyclin-dependent kinase inhibitors p16 and p19. p16 and p19 are involved in the inhibition of self-renewal and proliferation genes. Thus, activation of Bmi-1 promotes self-renewal.

hematopoietic and neural stem cells [113, 131]. This gene, known also as Pcgf4, is a member of the Polycomb family of proteins, a component of the PRC1 (Polycomb Repressor Complex 1) complex, and a chromatin modifying repressor [108]. Its role in self-renewal may be independent of PRC1 and is likely associated with its inhibition of the CDKN2A gene, which codes for two different proteins: Ink4a (p16) and Arf (either p19 or p14 in mice and humans respectively). Both of these proteins are cyclin-dependent kinase inhibitors and are responsible for the suppression of cell proliferation [129] (Figure 2.11). Activation of Bmi-1 is associated with the promotion of cell proliferation and self-renewal in normal hematopoietic stem cells [131], as well as leukemic stem cells (LSC) [106].

Another important shared pathway that plays a significant role in self-renewal is the Wnt/β-catenin pathway. The mechanism of action of this pathway was already described in Section 2.1.3. Wnt/β-catenin pathway activation occurs in many cancer types, including breast and brain cancer [143]. β-catenin accumulation has also been
observed in common myeloid leukemia in the GMP progenitor population when the
disease advances to a blast crisis [83, 108], which suggests the possibility of progenitor-
to-LSC transformation in these patients. Mutations in the degradation machinery (i.e.,
APC mutations) have been detected in many colon cancer types as well [143].

2.2.5 Metastasis and stem cells

Cancer is often associated with the development of metastatic tumors. Metas-
tasis is the process of the invasion of a new distant organ from the tumor growth of
origin. Understanding metastatic cancer has been a center of focus of a large body of
research. However, I will only present selected definitions, observations and studies in
this dissertation that relate directly or indirectly to stem cell biology. A nice overview
of metastasis is presented in an article by Bacac and Stamenkovic [10].

A primary tumor that successfully metastasizes to a new organ or tissue has to
complete several stages: separation from the tumor of origin through the degradation
of the basal membrane, migration and invasion of the stromal layers, transition into
the blood stream through a process called *intravasation*, survival in the blood stream,
transition out of it through another process called *extravasation*, preparation of the new
local microenvironment for the development of micrometastases and finally proliferation
and establishment in the new location [10].

The *epithelial-mesenchymal transition (EMT)* is often considered one of the
hallmarks of metastatic cancer, as it is indispensable to the successful invasion of a
new organ or tissue. This event refers to the process of conversion of cells from a more
epithelial-like state to a more mesenchymal-like state. Specifically, epithelial-like cells are well-organized, closely attached to each other, and are not generally considered migratory. Mesenchymal cells, on the other hand, are morphologically different (more spindle-shaped), are not well-organized and are more readily migratory. At the molecular level this transition can be observed in the change of the markers that the cells undergoing this transition express. For example, before the transition cells express E-cadherin. E-cadherin is associated with the maintenance of the epithelial cellular nature and exhibits properties of a tumor suppressor gene, as mutations in this gene drive the development of various metastatic carcinomas [32]. When cells undergo EMT, they lose the expression of E-cadherin and gain new markers, such as N-cadherin, fibronectin, vimentin and others [10, 105, 176].

In the past several years some number of studies have linked metastasis and stem cells through self-renewal and the EMT process. Interestingly, to successfully invade a new tissue, a tumor has to establish a stable colony at the distant site of metastasis, which means that as most cancer cells have an only limited proliferative ability, the cells that drive the metastatic growth expansion have to be able to actively self-renew.

One link between metastasis, poor cancer prognosis and stem cells was the identification of an 11-gene death-from-cancer signature, consisting of genes related to the Polycomb Repressor Complex (PRC) [63]. Specifically, this 11-gene signature was predictive of patients with poor survival rates, death after treatment and cancers with high metastatic potential. In the center of this signature was Bmi-1, which as already
mentioned is an important stem cell marker in hematopoietic, neural, intestinal and some cancer stem cells. The discovery of this signature and the involvement of the Bmi-1 pathway suggested an indirect role for stem cells in tumor metastasis. The theory that emerged from the study is that stem cells may be recruited to the tumor mass, where through the process of a rare fusion between the stem cell and a cancer cell, a new cell is formed that is neoplastic in nature, but has acquired additional stem cell-like properties which confer upon it the metastatic ability [62].

Several more recent studies also implicate self-renewal and EMT in metastasis, not through the Bmi-1 pathway, but as the result of activation of the canonical Wnt pathway. In particular, breast cancer metastasizes to lung show activation of Lrp6 (the co-receptor of the Frizzled receptor), where self-renewal is directly modulated by TCF/LEF transcription factor activation. Twist, an embryonic transcription factor, which is causally related to the epithelial-mesenchymal transition [199] is a downstream target of TCF/LEF. Inhibition of Lrp6 and the Wnt pathway reduces the self-renewal and metastatic capacity of these cancer cells and silences Twist and therefore the EMT modulation [45]. Wnt pathway activation and TCF/LEF have also been recently shown to regulate the self-renewal capacity of lung adenocarcinoma cells that metastasize to the brain [119].

There is now also some evidence for a direct link between EMT and stem cell-like properties, as shown in mammary carcinomas. Specifically, as mentioned previously Twist has the ability to induce the EMT transition of epithelial cells [87, 199]. If Twist is purposefully expressed in mammary epithelial cells when it should otherwise
be repressed, it causes their transition to a more mesenchymal morphological state. Besides the transition from epithelial to mesenchymal marker expression, these cells begin to also express the typical markers of normal and cancer mammary stem cells — CD24 and CD44 [4], as shown by FACS flow cytometry [110]. These transformed cells display not only the molecular, but also the functional features of mammary stem cells. The authors of the study show that the converse is true as well: normal mammary stem cells also express EMT-related markers [110].

2.3 Side populations

One central tool for the identification of stem-cell enriched populations from various tissues and organs is the isolation of side populations. This technique was first introduced in the hematopoietic system through the isolation of side populations from the bone marrow [163], but has since then been successfully used in many other tissue types, including liver [196], lung [168], brain [92], testis [53], and breast [193].

So what are side populations? They represent populations of cells with high efflux rates that actively remove any Hoechst 33342 or Rhodamine 123 dyes and exhibit low fluorescence levels after staining. In a typical FACS plot, they look like a collection of cells to the side of the fluorescence levels associated with most cells – the main population – which is how they received the name “side population.” A FACS plot toy example is shown in Figure 2.12 to illustrate this concept.

Even though the side population phenotype is not believed to be sufficiently
specific to be used as the sole marker of a stem cell, side populations are enriched for stem cells [33]. The abundance and enrichment vary between different tissues and organs, but within bone marrow the abundance ranges around 0.05—1% [33, 64]. Stem cell-enriched populations show the highest efflux activity and the rate of efflux is thought to be correlated with the level of commitment of cell populations, so it decreases as cells become more mature [65].

Different theories exist as to the cause of this unexpected phenotype. One theory suggests that because of their importance, stem cells must have a very active protein transporter system. The transporter system includes ATP-binding cassette (ABC) transporters that hydrolyze ATP to move various compounds across the cell membrane. ABC transporters include many proteins, among which the most well-known are Multi-Drug Resistance 1 (MDR1), ATP-binding cassette sub-family G member 2 (ABCG2), ATP-binding cassette sub-family B member 1 (ABCB1). They are heavily used by stem cells to remove various toxins and drugs and preserve the cell intact from potential harm [33, 42]. As a result of this ability, these cells can also remove dyes at a very fast rate.

2.4 High-throughput technologies and stem cells

2.4.1 High-throughput technologies

The development of various high-throughput technologies and methods over the last decade has enabled the accumulation of an enormous databank and has facil-
Hoechst Red 660/20
Hoechst Blue 424/44

side population (SP)
main population (MP)

Figure 2.12: A FACS plot for a toy example that illustrates the isolation and definition of a side population. Each axis represents Hoechst dye labeling. The cells that have efficiently removed the dye (the “side population”) are located in the lowest left hand-side corner of the quadrant.

...iterated many scientific discoveries. High-throughput technologies have been invaluable, because of their ability to measure thousands of quantities in a given sample within a short time frame and to create a quantitative profile that describes the state of the sample at the point of measurement.

In the next few paragraphs, I introduce very briefly some high-throughput technologies and high-throughput data types as they relate to the work in this dissertation.

2.4.1.1 Microarrays

Microarray technology has been one of the dominant technologies to emerge in the past 15 years. The purpose of microarrays is to capture the relative or absolute level
of expression of all transcripts in a given sample or in a pair of samples. Expression profiling, or the measurement of the level of mRNA produced from a given gene, is not a new concept and has been widely used in molecular biology since the 1970s. However, the experimental scale has grown immensely and within the last five years microarray platforms have reached genome-wide scale.

Conceptually, microarray technology can be described as follows: you create a chip with thousands of features (currently with genome-wide coverage). Each feature represents a probe or a short RNA sequence attached to the chip and complementary to some mRNA transcribed from some gene of interest. You select a sample of interest, or in the case of some microarray technologies, two samples of interest. The mRNAs in the sample are going to hybridize and attach to their complementary probes on the microarray chip. In the case of two samples, the mRNAs from the individual samples competitively hybridize to the array, based on the abundance of the specific mRNA in each sample. The mRNAs are fluorescently labeled so high abundance of a given mRNA will spot brightly, while low mRNA abundance will not and both can be assessed quantitatively [46].

Two general types of microarrays are in widespread use: oligonucleotide arrays and cDNA spotted arrays. Oligo arrays consist of short probes corresponding to small segments of genes or other genomic elements. The probes can vary in size from 25 to 60nt depending on the microarray design (Affymetrix and Agilent). The platforms can also vary and range from the silicon chip (Affymetrix technology) to micro-beads (Illumina technology). cDNA spotted arrays use larger pre-assembled probes spotted
directly onto a glass slide [46]. Spotted arrays are often used in the context of competitive hybridization, where two samples are introduced instead of one and the resulting measurements represent the relative abundance of the mRNA in one of the samples versus the other. Usually, when the relative abundance of mRNA between two samples is of higher interest than independent absolute measurements, this context is preferred as it avoids noise issues associated with the non-reproducibility of individual spot measurements.

2.4.1.2 SAGE

Serial analysis of gene expression (SAGE) technology gained prominence in the 1990s and is similar to the microarray technology in that it is used to measure mRNA levels in a given sample. Unlike microarrays in which a probe must be present on the chip to be measured, SAGE is a sequencing-based technology and does not require knowledge of all the molecules present in the sample. The general technique can be described as follows: mRNAs are isolated from the sample of interest and reverse-transcribed into cDNA. Then short sequence tags sufficiently long to identify transcripts uniquely are isolated from individual mRNAs and ligated together. The ligated sequences are then amplified, sequenced and analyzed. The abundance of a given transcript is measured as the count of the number of that transcript sequenced from the sample [46].
2.4.1.3 RNA-Seq

RNA-Seq is the most recently emerged technology and considered as the transcriptome technology of the future, perhaps soon to replace microarrays. Similarly to SAGE, because it is sequenced-based, the RNA-Seq technique does not require prior knowledge of the mRNAs potentially present in a given sample, which may be an advantage over microarrays. Though this technique is quite new, some stem cell studies that use RNA-Seq have already been published in the field [38] and many more are expected to emerge in the next several years.

2.4.2 High-throughput data

The above three technologies are some of the main high-throughput techniques used to generate expression data in stem cell experiments. Because expression data is the core of my stemness analysis, I next review the most common methods used to identify differentially expressed genes. I take the opportunity to introduce some other high-throughput non-expression data types, though description is restricted to the protein interaction data used in my stemness research. I also present some of the main public repositories for these data, as they are important for data collection.

2.4.2.1 Differential expression

Expression data are most commonly used in a comparative manner, where mRNA levels are compared in a pairwise fashion between two samples of interest. Depending on the experimental platform, these samples can include
• normal tissue vs. normal tissue, e.g. embryonic stem cells vs. neural stem cells

• normal tissue vs. cancer tissue, e.g. normal liver cells vs. liver cancer cells

• time-course experiments, e.g. undifferentiated cell at 0h vs. differentiating cell at 96h after treatment with a differentiation-inducing drug

• normal tissue vs. universal (and/or pooled) reference, e.g. liver stem cells vs. universal mouse reference

Genes that are highly expressed in one tissue, but not the other are often most interesting. Differential regulation of a gene is expected to be associated with the specific functional roles the gene may play in the tissue, where it is found to be differentially expressed. Many techniques can be used to identify differentially expressed genes. The two most common methods used in the stem cell literature are SAM [181] and the fold-change [40] methods.

Based on much of the published stem cell data in the last decade the fold-change method is perhaps the most utilized technique. For a given probe or gene of interest $G$, the fold change between samples $S_1$ and $S_2$ is measured as the ratio between the (normalized) expression $E$ of the gene in each sample:

$$ FC(G) = |\log\left(\frac{E_{S_1}}{E_{S_2}}\right)| > c. $$

If the fold-change $FC(G)$ exceeds the cutoff $c$ and $E_{S_1} > E_{S_2}$, the gene is said to be differentially upregulated in $S_1$. Similarly, if $E_{S_1} < E_{S_2}$, the gene is marked as
differentially downregulated in $S_1$. Common fold-change cutoffs range between 2- and 3-fold.

The other commonly used method for the identification of differentially expressed genes is *Significance Analysis of Microarrays (SAM)* [181]. The differential expression statistic or the *d-score* reported by SAM for each gene can be viewed as a modified gene-specific t-statistic most often measured between two classes, samples, or conditions. The significance of the d-score is evaluated through permutation analysis and significantly differentially expressed (upregulated or downregulated) genes are usually selected based on a FDR cutoff.

Both of these methods provide as output a list of genes differentially upregulated or downregulated between two classes or samples. I will come back to these lists in the next chapter as they are a particular important component of meta-analysis.

2.4.2.2 Protein-protein interactions

The interaction between proteins is often central to the exercise of their functional roles in the cell. Many protein complexes require the assembly of individual protein components before the complex can become functionally active. Therefore, studying protein-protein interactions on a wide scale has been of biological interest to many researchers. Protein-protein interactions are of particular interest to the study of stemness as they could be used to test for putative protein complexes with central roles in stem cells.

Various high-throughput methods have been developed not only for the mea-
surement of mRNA levels in a given cell in a given condition, but also for the measurement of protein levels, as well as the interaction between proteins in a given cell or condition. Such methods include yeast two-hybrid, co-immunoprecipitation, tandem-affinity purification and many other techniques. These technologies produce interaction data on a high-throughput scale and even though it may be impossible to ascertain the specificity of the inferred protein interactions, we still get a glimpse of the global protein state of the cell.

2.4.3 Public data repositories

There are many public data repositories available for use to the larger research community. The most central databases for expression data include Gene Expression Omnibus (GEO) [51] and Array Express [132]. Other gene expression databases used in the past include the Stanford Microarray Database (SMD) [43], however the import of all publicly available data in SMD into GEO has diminished the need for independent usage.

Public repositories for protein-protein interaction data include BioGRID [166], Bind [11], Dip [197], and HPRD [134]. These databases contain both interactions characterized by the techniques mentioned in the previous section, as well as computationally predicted interactions. Large-scale experimentally verified protein complex data in mammalian species is also available in the CORUM database with independently verified mouse, rat and human protein complexes [148].
2.5 Summary

In this chapter, I introduced the core of the biological background needed to understand the basis for my stemness analysis, as well as the results generated by it. Section 2.1 gave some basic stem cell definitions and described the key properties of normal stem cells. It also gave a broad overview of the most common normal stem cell types studied in the literature. It should be noted that many more stem cell types do exist and the description is by no means exhaustive. Section 2.1.3 introduced some known pathways associated with self-renewal. Section 2.2 described tumor heterogeneity, the different cancer evolution theories that have evolved to explain it, and the possible links between normal stem cells, cancer stem cells, and metastasis. Section 2.3 described the stem-cell enriched side populations, while Section 2.4 introduced the main gene expression high-throughput technologies, the public repositories used for such data, and the main differential expression inference methods.
Chapter 3

Meta-analysis

This chapter aims to introduce meta-analysis and describe some of the most commonly used techniques in this field to provide the context for the methodology used in my stemness work. Section 3.1 provides a brief overview of the meta-analysis field and what the concept of *meta-analysis* actually means and entails. Section 3.2 briefly discusses how meta-analysis can be used for the analysis of microarray experiments. The most commonly used integration techniques are introduced in Section 3.3, while the final Section 3.4 analyzes the applicability of the various integration techniques to my stemness research and justifies my final technique choice.

3.1 Overview of Meta-analysis

Meta-analysis entails the statistical evaluation of a large number of studies for the measurement of the size and significance of a certain effect [20]. It is common to consider the size of an effect measured by a given study. This effect size could represent
• the extent of a drug effect on a set of patients,

• the level of upregulation of a set of marker genes in a given cell population, or

• the odds of survival after metastasis of a specific tumor type.

The only requirement is that the effect size has to be a measurable quantity. So, how does meta-analysis improve upon previous effect evaluation methods? In measuring one combined effect size from many data sources, meta-analysis can take into account many important, but indirectly relevant factors, such as the size of each study, the variability of each experiment and its results, the methodological differences between individual studies, etc.

3.2 Meta-analysis and microarray data

Robust cross-study statistical analysis has become very important with the emergence of gene expression high-throughput technologies. Meta-analysis lends itself particularly relevant to microarray analysis, because it provides a consistent framework for the integration of information across different studies explicitly accounting for differences in platforms, methodology, cell types and other technical factors that can affect the inference of a true biological effect.

Many different meta-analysis methods have been applied to microarray data [36, 66, 104, 160, 167, 200]. These can be broadly classified in three categories: differential expression-based, co-expression-based and most recently, differential co-expression-based meta-analysis. I will not provide a separate overview of the results and methods
used by independent studies, as most of the recommended and used techniques are summarized in Section 3.3. The most relevant individual studies will be mentioned in the next chapter, which specifically focuses on previous gene expression-based approaches to stemness.

Because of the recent development of meta-analysis, books and review papers have just begun outlining general guidelines – almost in a cookbook recipe style – for the application of meta-analysis to microarray data. A great example of a review study that presents a set of potential microarray meta-analysis guidelines is the recent study of Ramasamy, et al [142]. Even though I do not explicitly follow all recommendations outlined in this review, I provide a summary of the recommended steps and techniques as they closely reflect the steps I followed in the development of the methodology for my stemness studies. Any deviations from the outlined recommendations are discussed where necessary.

3.3 Techniques for combining study-specific effects

One important aspect in the meta-analysis process involves the selection of appropriate studies, along with the pre-processing of the input data [142]. The pre-processing steps can include preparation of the data associated with each study, such as normalization, or mapping of all probes to a common gene space. However, once the pre-processing has been completed, the most crucial elements of meta-analysis are the inference of a study-specific effect and the combination of the individual study-specific
The choice of appropriate techniques for analysis often depends on the input data type. Most commonly, researchers use either the raw data directly, or alternatively the sets of differentially expressed genes that can be estimated from the raw data. For both of these input types, a measurement can be generated for every probe in every sample.

Several meta-analysis techniques are most commonly used. The two most common methods for the integration of effects across multiple studies include p-value or rank combination methods [142]. However, as they are not as relevant to my stemness analysis, I do not discuss them in detail in this dissertation. Other meta-analysis integration methods include vote counting and effect size combination.

### 3.3.1 Vote counting

One of the simpler, but often very useful meta-analysis approaches is vote counting. The researcher can simply count the number of studies in which a gene is marked as present. To assess if the count is actually significant, one can employ permutation techniques to build a null distribution and estimate a significance score. Another common strategy used to evaluate the significance of overlap is to compare the amount of observed overlap to the expected amount of overlap based on the binomial distribution.
3.3.2 Effect size combination

The final meta-analysis technique incorporates indirectly relevant factors to the measure of one common combined effect size. Meta-analysis theory separates fixed and random effects. The fixed effects model assumes that there is one true common effect size and all of the variability observed in the measurement of individual study effect sizes stems from within-study measurement error [20]. On the other hand, the random effects model assumes that there is a distribution of true effect sizes from which the individual study effect sizes are drawn and the observed variability can be modeled as a combination of both the between-study variability in effect sizes, as well as the sampling error [20].

In either case, however, the most important distinguishing feature of this approach is the use of the inverse variance approach. What this generally means is that after the measurement of an individual study effect, each study’s contribution to the overall combined effect is weighted by the inverse variance associated with that study. The only difference between the fixed and random effects models is that the variance in the former case involves only the within-study variance, while in the latter case the inverse variance weight has two components: the within- and the between-study variance.
3.4 Choice of techniques for the stemness meta-analysis

The techniques described in the previous section have been applied to different studies and data types, but there is no universal consensus on the best input data choice or the best method for combining measurements across studies.

Ramasamy et al. [142] discusses the advantages of using the raw data, as studies can use different methodologies to infer significant gene lists of differentially expressed genes. For example, to identify significantly upregulated genes in liver, study A may use a fold change method with a cutoff of 2, while study B may use a cutoff of 4. The difference in cutoffs may place these two studies on unequal grounds when assessing the combined set of significantly upregulated genes.

However, even though the use of raw data can be preferable or advisable when combining results across studies performed on the same microarray platform within similar tissues, the use of pre-defined lists of differentially expressed genes is invaluable when combining effects across different microarray platforms, precisely because it avoids the need for direct cross-platform integration. As to the differences in the methodology used by different studies for assessment of significant genes, many authors seem to prefer similar methodology with closely related cutoffs for the inference of significantly differentially regulated genes. Also, one can explicitly model these differences in a manner similar to the inverse variance approach method by accounting for the level of specificity or stringency in the measurement of differentially expressed genes in each study. I describe my implementation of such an approach for the stemness meta-analysis.
As to the techniques outlined by Ramasamy [142], some are less relevant to the study of stemness, especially if one uses the lists of differentially expressed genes from each study as input. Both the rank combination and p-value combination methods require an explicit assignment of significance to each differentially expressed gene in every study, which is not easily available (p-values) or even meaningful (ranks) for many studies. Because of that, I use a combined approach between a vote-counting and an effect-size technique to define a recurrence scoring method, which can identify reproducibly upregulated genes across many different stem cell studies.

3.5 Summary

The goal of this chapter was to introduce meta-analysis and some of the basic techniques used to infer combined effect sizes across many different studies. Section 3.1 gave a brief history of the evolution of meta-analysis and its basic premise. Section 3.2 introduced the use of meta-analysis in microarray studies. Section 3.3 outlined the four most common techniques used to combine study-specific effect sizes. Finally, the last section 3.4 discussed some of the pros and cons of using individual techniques on stem cell data. This discussion is further extended in the next chapter, which introduces some of the pre-existing methods and applications of meta-analysis to the search of a common molecular program shared by all stem cell types.
Chapter 4

Previous expression-based approaches to stemness

The goal of this chapter is to give a broad introduction to the previous work in the literature on stemness. The previous research can be broadly divided into two categories: gene-level approaches and global-level approaches. Section 4.1 introduces the most relevant gene-level approaches, including the three main studies that led the discussion in this field. Section 4.2 discusses the more global approaches to the study of stemness and introduces the closest study in the literature to my own stemness work. It also outlines the elements still missing in the existing literature and highlights the points of contribution of my work.
4.1 Gene-level approaches to stemness

4.1.1 Founder studies

The concept of stemness, or a shared molecular program between stem cells, gained prominence with the introduction of the high-throughput microarray technology, which made possible the large scale analysis of gene expression. Researchers hypothesized that as functionally defined cells, stem cells with their two characteristic properties of self-renewal and multi- or pluripotency may share a common molecular program, i.e. there may be specific genes that allow all stem cells to retain their stem cell state.

The first high-throughput gene expression experiments on stem cells were published in 2000–2001, but the most relevant studies to stemness are three foundation experiments published in 2002 and 2003 by Ivanova et al. [79], Ramalho-Santos et al. [141], and Fortunel et al. [56]. Each of these studies examined several stem cell types and identified a set of differentially upregulated and differentially downregulated genes between each stem cell type and a corresponding differentiated or mixed cell type. Both Ivanova and Ramalho-Santos studied embryonic, hematopoietic and neural stem cells, while Fortunel examined embryonic, neural and retinal stem cells [56, 79, 141]. Notably, all three studies used a similar microarray platform, but the downstream comparison populations used to identify differentially expressed genes in each stem cell experiment varied between each study. Despite some of the methodological differences in stem cell and comparison cell definitions, if a common molecular program did exist, it should have emerged from the three study comparisons, especially since the comparisons within a
stem cell type between the different studies identified many common differentially upregulated genes, so the differences were not sufficient to mask the common mechanisms within a single stem cell type.

Interestingly, each study could also identify a high number of genes commonly upregulated across stem cell types within the study (each Venn bubble in Figure 4.1), but surprisingly as Fortunel et al [56] reported, there was only one gene commonly upregulated across all three studies and nine stem cell populations (Figure 4.1). The

![Venn diagram](image)

Figure 4.1: Overlap between the genes upregulated in four stem cell types in three stemness founder studies. Each Venn bubble represents all genes that are commonly upregulated between all tested stem cell populations within that study. Figure is adapted from the original study of Fortunel et al. [56].

Fortunel study used a simple vote-counting procedure, which is highly appropriate for meta-analysis on a small number of studies [142], and directly counted the number of experiments in which each gene was identified as significantly upregulated. Integrin α6

54
(Itga6) was the only gene that shared upregulation across all stem cell types and studies. Since then, other transcriptional profiling studies have also shown a low overlap with the genes selected by the founder studies and few stemness genes have been identified to date [73].

There are several possible factors that can account for the lack of stemness genes [50]. The first one, as mentioned above, involves methodological differences between the comparison populations, but others include

- Absence of stemness genes from microarray platform
- Transient expression of stemness genes that has not been captured by a “static” population
- Expression of stemness genes in differentiated cells with regulation at the post-transcriptional level
- Stemness defined at a modular level, such that individual genes are dispensable, but shared pathways, protein complexes, or functional gene module are necessary for the maintenance of stem cell state.

While some of the more technical factors have since lost their relevance (for example, most recent experiments have been performed on genome-wide microarray platforms) and others cannot be addressed with the use of microarrays and gene expression (for example, if the stemness gene levels are determined post-transcriptionally), more evidence is pointing the search for stemness to more global approaches, such as the identification of pathways and complexes shared between all stem cell types.
Before I discuss the global-level approaches to stemness, I first focus on another
gene-based study with a similar aim that identified stem cell markers through data
integration. Krzyzanowski et al. [102] identified genes that distinguish between two
conditions, such as a stem cell population and a selected control condition. Specifically,
they analyzed every probe set (gene) to find genes that create “gaps” between samples
in their data, i.e. there is a partition of the data under which the probe set showed
high level of expression in one of the subpartitions and a low level of expression in the
other. Each probe set was scored for its expression in the two sample partitions using
the U-statistic associated with the Mann-Whitney non-parametric test [102].

The approach of Krzyzanowski et al. [102] is relatively similar to other ap-
proaches that aim to directly identify consistently differentially expressed genes in a large
set of studies (a vote counting strategy). The difference in the Krzyzanowski study is
that the sample partition is not predefined, so it is possible to discover marker genes
associated with different stem cell types. Their analysis identified 426 possible stem cell
markers for many different individual stem cell types, but many of these markers have
already been discovered.

The authors also explored the evolution of stem cell marker genes by observing
superfamilies of proteins that have been independently suggested as stem cell markers
but show very different tissue-stem cell specificities. Even though the complementarity
of protein function and specialization to different stem cell types is hinted at in this
study, the authors choose the simplest possible data integration method and each study
in a given cell type was treated as a replicate. Such treatment, however, was only

56
possible because the same microarray platform was used in all datasets included in the study.

4.2 Global-level approaches to stemness

Recently, researchers have started looking at more global mechanisms to understand stemness and several independent efforts have been made in this direction. One common feature of all global-level stemness studies is their use of pathways and functional gene sets to either distinguish between different stem cell types, or find common mechanisms.

A large-scale effort by Kluger et al. [94] attempted to identify gene set modules that could classify between different stem cell populations [94]. The study focused on a number of stem, progenitor, and differentiated cell populations, such as neutrophils, monocyte, macrophages, lymphocytes, and hematopoietic stem cells and the expression profiles all populations were assessed using oligonucleotide arrays. To find large gene sets that could differentiate between these populations, the authors applied principle component analysis (PCA) to various pathways and functional gene-sets as defined by Biocarta, KEGG [85, 86] and Gene Ontology [7]. They identified gene sets that could be used as good features in a classification setting because of their ability to separate accurately the different lineages [94]. These results suggested that gene sets and pathways could play a significant role in classification, due to their ability to capture more subtle changes in transcription that may be otherwise missed by differential expression.
and gene-level approaches.

Another study with a similar aim was recently published by Doherty et al [47]. The goal of the study was to identify higher patterns of expression shared between different stem cells as compared to transit-amplifying and differentiated cells, based on GO gene set activation differences between cell types at different levels of differentiation. One interesting aspect of this paper was that the authors avoided the use of raw data and cross-platform analysis by reducing the raw data for each sample to a single vector, where each entry was associated with one GO category. The specific value associated with the entry corresponded to the fraction of genes in that category that were activated in the sample of interest. Finally, a two-tailed t-test was applied for each GO category to identify categories that could significantly distinguish between stem, transit-amplifying and differentiated cells.

In 2008, Muller et al. [117] introduced the PluriNet - a protein-protein interaction network of embryonic-stem cell-specific genes that can be used to define and classify the embryonic stem cell state. To achieve this goal, the researchers applied non-negative matrix factorization to their stem cell data, which allowed the clustering of experiments with similar expression patterns, including embryonic and induced pluripotent stem cells.

Perhaps the most relevant work to the study of stemness is the “stem cell module map.” Wong et al [194] initially tried to construct a stemness signature in mouse and human through measurement of the activation of functional gene modules in each stem cell and differentiated cell type used in their study. Even though the authors
did not find a shared stemness signature, they defined two independent signatures: an embryonic-stem cell-like and an adult stem cell-like signature and showed that one of the key regulators of the embryonic stem (ES) cell-like signature is c-myc \cite{194}. This transcription factor has the ability to induce the ES-signature in a normal epithelial cell, causing its transformation to an epithelial cancer stem cell. Biologically, this regulatory dependency allows the induction of a cancer-like state and the potential creation of pluripotent cancer stem cells, which could be invaluable for more advanced functional laboratory studies.

The advantage of these global-level approaches to stemness is the recognition of the stochastic nature of gene expression and the dispensability of individual genes for the maintenance of the stem cell state. However, one aspect that all global-level approaches to date overlook is the potential functional redundancy afforded by gene duplications and paralogs. The only study to hint at functional redundancy between evolutionarily-related proteins is the study of Krzyzanowski et al \cite{102}, but their focus is still primarily gene-based and the study relies on very simple data integration techniques. Another disadvantage of the majority of previous studies is the lack of a unified scheme for integration of stem cell experiments that could account for the differences in specificity, the potential underlying correlation between individual experiments, and other factors that can potentially skew the analysis, such as the influence of primary versus cultured cells.

The methodology I introduce in the next chapter tries to address these important issues and fill the voids in the current literature. I make use of meta-analysis
techniques to define a unified scheme for the measurement of reproducible and recurrent gene or module expression across many experiments. I also explicitly test the functional redundancy hypothesis, evaluate the contribution of homolog-based gene modules to the maintenance of the stem cell state and compare their role in stemness to the one of functional gene modules.

4.3 Summary

This chapter presented a summary of the previous work in the literature on shared stem cell mechanisms. Section 4.1 introduced the founder studies – the best known gene-level stemness approaches, as well as other gene-level analyses. Section 4.2 briefly described the more global approaches in the literature used to either distinguish between different stem cell types, or between cells at various stages of differentiation. This section also described some of the possible improvements that my stemness meta-analysis methodology (introduced next) could contribute to the literature.
Chapter 5

Stemness Meta-Analysis Method

This chapter provides a description of the methodology used in this study to identify and measure stemness mechanisms in different types of stem cells. Section 5.1 gives a brief outline of the Stemness Meta-Analysis method. The method that I describe has two major inputs: Section 5.2 summarizes the input profiling data used to generate mouse and human stem cell compendia, while Section 5.3 describes the generation of functional and evolutionarily-related gene modules. The subsequent three sections describe in detail the individual steps—recurrence (Section 5.4), diversity (Section 5.5) and specificity (Section 5.6)—of the meta-analysis and the selection of appropriate scoring methods for each step. Section 5.7 uses these three scoring measures to classify modules into different pattern types, including the patterns associated with stemness. To aid the understanding of the reader, a graphical flowchart overview of the SMA method is shown in Figure 5.1 and it summarizes the input, steps and classification patterns described in the chapter. Section 5.8 outlines the two types of stemness-associated
Figure 5.1: Overview of stemness meta-analysis method. The two main inputs to the method are gene modules and a large gene expression compendium. The gene modules (Input A) encompass both evolutionarily-related sets of genes, and functionally-associated sets of genes, such as pathways and protein complexes. The large stem cell gene expression compendium (Input B) consists of lists of differentially upregulated and downregulated genes (PGLs) derived from the literature. The method identifies recurrently expressed modules from the input data (Step 1), and classifies their expression pattern based on two additional scores – cell diversity and specificity. The final output consists of stemness “on” modules (red dashed line), which are upregulated across most stem cell types, and stemness “off” modules (green dashed line), which are downregulated across most stem cell types.

modules – stemness “on” and stemness “off” modules. Finally, Section 5.9 describes a few scoring metrics that could be used to define a stemness index score – a score that measures how stem cell-like a gene signature is.

5.1 Stemness Meta-Analysis method overview

The Stemness Meta-Analysis (SMA) method uses two input types. The first input consists of evolutionary or functional gene modules. Here the term module is used in an all-inclusive manner and incorporates individual genes as well. The second input consists of the published gene lists (PGL) of differentially expressed genes as identified
by individual stem-cell-related studies. This second input can also be visualized as a matrix, where each row-column entry can be represented by one of three values: differentially upregulated, not differentially upregulated, or not tested (Figure 5.2).

Two such matrices are used: one is associated with the stem cell data, and the other is associated with the differentiated cell data, such that the genes that are differentially upregulated in the differentiated cells can also be viewed as the differentially downregulated genes in the stem cells. The stem cell upregulated and downregulated gene lists are separated for both ease of computation and clear determination of the role (“on” or “off”) of individual gene modules.

Let the set of genes tested in experiment $j$ be denoted as $W_j$. I will define the set of genes upregulated in experiment $j$ as $U_j$ and the set of genes downregulated in experiment $j$ as $D_j$. I will further introduce two indicator matrices $X$ (an upregulation status matrix) and $T$ (a test status matrix), such that $x_{ij}$ is 1 if the gene $i$ is upregulated in experiment $j$, and it is 0 otherwise. Similarly, $t_{ij}$ is 1 if the gene $i$ is tested in experiment $j$, and it is 0 otherwise.

An obvious advantage of using PGLs reported by the authors of each original study is that it allows the results from different microarray platforms to be compared. The set of reference cell populations can vary between different studies and thus, the genes identified as upregulated (or downregulated) in stem cells from different studies can also vary. However, the goal of the analysis was to detect robust and reproducible stem cell-specific expression over multiple studies, so while the variability might impact the detection sensitivity, I expect few false positives.
Figure 5.2: Published gene list (PGL) input form to Stemness Meta-Analysis Method (SMA). The data can be represented as a trinary matrix, such that each column corresponds to a single literature-derived experiment, while each row represents a single gene. In a given row and column, the entry corresponds to one of three marks: “upregulated” (green star), “not upregulated” (red cross), or “not tested” (white square).

The SMA method follows several steps: the initial step tests for the existence of genes and gene modules with stem-cell-associated patterns of expression across multiple studies. To measure whether a module (or gene) is upregulated in a significant number of studies, I compute a recurrence score, which builds upon previous meta-analysis techniques and incorporates the redundancy among studies. Some PGLs have a high degree of overlap with other PGLs in the input data, such as lists of differentially expressed genes in the same stem cell type or those obtained from the same study. To avoid double-counting redundant information, I group highly similar PGLs into equivalence classes and weight the score such that each equivalence class is allowed to contribute one unit of weight.

The method then applies a cell-diversity measure, based on information-theoretic entropy, to quantify how a module’s (or gene’s) upregulation is distributed across different stem cell types. Genes or modules upregulated to the same extent in each stem cell
type (i.e. same fraction of studies for each type) are associated with a high cell-diversity, while those expressed disproportionately are assigned a low cell-diversity.

Finally, to measure whether a module’s (or gene’s) pattern of upregulation is specific to stem cells, I also measure the level of its upregulation in differentiated cells. Genes or modules found to be significantly upregulated in stem cells may also have roles in differentiated cells. However, those with upregulation specific to stem cells may shed light on stemness properties. I quantify the degree to which a gene was not upregulated in differentiated cells using a specificity score, based on the level of recurrence of the module (or gene) across differentiated cells.

Altogether, modules that exhibit significant recurrence scores across stem cell experiments, significant cell-diversity across most stem cell types, and significant specificity to stem cells are labeled as stemness “on” modules. However, the SMA method can also be applied to the differentiated cell data as well, so modules that show significant recurrence scores across differentiated cell experiments, significant cell-diversity across most differentiated cell types, and significant specificity to the differentiated cells are identified as stemness “off” (differentiation) modules. The stemness “on” modules include genes that need to be consistently upregulated by stem cells, while stemness “off” modules include genes that need to be consistently downregulated by stem cells (or upregulated by differentiated cells). To avoid redundancy, the descriptions of the methodology focus on the selection of stemness “on” modules, but the selection of stemness “off” modules is exactly the same, only based on the differentiated cell data.

I use the recurrence, diversity and specificity scores to test for many module-
level patterns of upregulation. A gene may be weakly associated with stemness (e.g., differentially expressed in a fraction of studies) because of the stochastic nature of gene expression. The cell’s use of alternate pathways and the ability of genes to compensate for one another, as evidenced by extensive genetic synthetic-lethal maps in several model organisms [27, 177], suggest that an examination of evolutionarily-related gene modules could identify stemness patterns. Genes also interact with one another in signaling cascades, complexes, and metabolic reaction pathways, so interpreting gene expression data using sets of genes that approximate modules of activity, rather than single genes in isolation, may also elucidate significant patterns of expression associated with stemness. A conceptual illustration of stemness at the module level is shown in Figure 5.3.

A module may have several possible significant patterns of upregulation in stem cells consistent with it playing a key role in stem cell function. I use the diversity score, coupled with recurrence and specificity, to categorize patterns of module upregulation. A module may contain genes upregulated across a diverse set of stem cells, which I quantify using the cell-diversity score introduced earlier. In addition, a module may employ many or a few constituent members, which I quantify with an analogous gene-diversity score. Combinations of cell-diversity, gene-diversity, and specificity represent possible module-level patterns as illustrated in Figure 5.4. All-for-all (AFA) modules express many members in most stem cell types and offer the clearest example of stemness-associated groups of genes (Figure 5.4: first row). One-for-all (OFA) modules express predominantly a single gene in the majority of stem cell types, even though they may have other gene members occasionally active as well (Figure 5.4: second row). All-
Figure 5.3: Module-level view of stemness. A stemness module (dashed line) could represent an evolutionarily-related set of genes $(a_1, a_2, a_3)$, a pathway $(b_1, b_2, b_3)$, or a protein complex $(c_1, c_2, c_3, \text{and } c_4)$, in which individual genes are not upregulated in all stem cell types, but the module through the cooperative upregulation (red) of its members is represented in all stem cell types.
For one (AFO) modules express many genes in a limited set of cell types and may represent modules with redundant gene functions important for a specific stem cell lineage but dispensable for others (Figure 5.4: third row). Constitutive module sets (CM) express many genes in most stem cell types, but also across a significant proportion of differentiated cell types (Figure 5.4: fourth row). Finally, constitutive gene sets (CG) express predominantly a single gene in most stem cell types, but also show significant upregulation in differentiated cells (Figure 5.4: fifth row).

5.2 Input data sources

5.2.1 Mouse profiling studies

5.2.1.1 Definitions and sources

I used 30 different studies, corresponding to 49 different mouse cell populations or 12 different stem cell types from published transcriptional profiling studies. The descriptions of the publications, cell types and other additional information on each study have been included in Tables 5.1 (primary/freshly isolated cell data) and 5.2 (cultured cell data). The experiments used as input to the analysis have been separated based on the source of the cells – primary or cultured. Primary cells are freshly isolated cells derived directly from mice tissues, while cultured cells are grown in vitro.

For each data source, I obtained the lists of upregulated clones in each stem cell population either directly from the publication, or through correspondence with the authors of the publication. In both cases, the lists were used either as directly reported
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Cell diversity</th>
<th>Gene diversity</th>
<th>Specificity</th>
<th>Pattern example</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-for-all (AFA) &quot;stemness&quot;</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>SC1 SC2 SC3 SC4 OR g1 g2 g3 g4</td>
</tr>
<tr>
<td>One-for-all (OFA) &quot;stemness&quot;</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>SC1 SC2 SC3 SC4</td>
</tr>
<tr>
<td>All-for-one (AFO)</td>
<td>-</td>
<td>+</td>
<td>+/-</td>
<td>SC1 SC2 SC3 SC4</td>
</tr>
<tr>
<td>Constitutive module (CM)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>SC1 SC2 SC3 SC4</td>
</tr>
<tr>
<td>Constitutive gene (CG)</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>SC1 SC2 SC3 SC4</td>
</tr>
</tbody>
</table>

Figure 5.4: Classification of modules based on cell-diversity, specificity and gene-diversity scores. The first column gives the pattern type names. The second, third and fourth columns indicate the criteria modules needs to pass to be associated with a given pattern type. A ‘+’ symbol that the score is above the threshold, while a ‘-’ suggests that the score is below the threshold. The symbol ‘+/−’ is used when the threshold is not important. The fifth column shows examples of each pattern type. Red indicates that a given gene is upregulated in the stem cell, while green indicates the gene is upregulated in the differentiated cell.
in the publication, or as inferred based on the methodology described in the paper.

After the initial selection of upregulated clones from each study, each clone was mapped to its corresponding Entrez Gene ID. If Entrez Gene did not have a mapping for a clone, the clone was excluded from further analysis. The final number of upregulated genes in each stem cell population can be found in Tables 5.1 and 5.2.

<table>
<thead>
<tr>
<th>Author</th>
<th>Class</th>
<th>Primary stem cell</th>
<th># upregulated</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ivanova</td>
<td>Hematopoietic</td>
<td>HSC</td>
<td>794</td>
<td>[79]</td>
</tr>
<tr>
<td>R-Santos</td>
<td>Hematopoietic</td>
<td>HSC</td>
<td>1476</td>
<td>[141]</td>
</tr>
<tr>
<td>Forsberg</td>
<td>Hematopoietic</td>
<td>LT-HSC</td>
<td>402</td>
<td>[55]</td>
</tr>
<tr>
<td>Akashi</td>
<td>Hematopoietic</td>
<td>LT-HSC</td>
<td>702</td>
<td>[3]</td>
</tr>
<tr>
<td>Kiel</td>
<td>Hematopoietic</td>
<td>FL-HSC</td>
<td>346</td>
<td>[90]</td>
</tr>
<tr>
<td>Kiel</td>
<td>Hematopoietic</td>
<td>FS-HSC</td>
<td>236</td>
<td>[90]</td>
</tr>
<tr>
<td>Kiel</td>
<td>Hematopoietic</td>
<td>Femur-HSC</td>
<td>401</td>
<td>[90]</td>
</tr>
<tr>
<td>Kiel</td>
<td>Hematopoietic</td>
<td>Pelvis-HSC</td>
<td>303</td>
<td>[90]</td>
</tr>
<tr>
<td>Kiel</td>
<td>Hematopoietic</td>
<td>Sternal-HSC</td>
<td>316</td>
<td>[90]</td>
</tr>
<tr>
<td>Kiel</td>
<td>Hematopoietic</td>
<td>HSC</td>
<td>928</td>
<td>[91]</td>
</tr>
<tr>
<td>Terslikh</td>
<td>Hematopoietic</td>
<td>HSC</td>
<td>30</td>
<td>[175]</td>
</tr>
<tr>
<td>Chambers</td>
<td>Hematopoietic</td>
<td>HSC</td>
<td>272</td>
<td>[34]</td>
</tr>
<tr>
<td>Fortunel</td>
<td>Retinal</td>
<td>RPC</td>
<td>1716</td>
<td>[56]</td>
</tr>
<tr>
<td>Stappenback</td>
<td>Intestinal</td>
<td>SiEP</td>
<td>159</td>
<td>[165]</td>
</tr>
<tr>
<td>Giannakis</td>
<td>Intestinal</td>
<td>SiEP</td>
<td>1440</td>
<td>[61]</td>
</tr>
<tr>
<td>Fevman</td>
<td>Intestinal</td>
<td>InSC</td>
<td>427</td>
<td>[54]</td>
</tr>
<tr>
<td>Mills</td>
<td>Gastric</td>
<td>GEP</td>
<td>135</td>
<td>[112]</td>
</tr>
<tr>
<td>Giannakis</td>
<td>Gastric</td>
<td>GEP</td>
<td>1366</td>
<td>[61]</td>
</tr>
<tr>
<td>Morris</td>
<td>Epithelial</td>
<td>EpSC</td>
<td>92</td>
<td>[114]</td>
</tr>
<tr>
<td>Tumbar</td>
<td>Epithelial</td>
<td>EpSC</td>
<td>106</td>
<td>[180]</td>
</tr>
<tr>
<td>Orwig</td>
<td>Spermatogonial</td>
<td>SSC</td>
<td>320</td>
<td>[126]</td>
</tr>
<tr>
<td>Kokkinaki</td>
<td>Spermatogonial</td>
<td>SSC</td>
<td>341</td>
<td>[96]</td>
</tr>
</tbody>
</table>

Table 5.1: List of non-cultured (primary) mouse stem cell profiling studies used in the mouse stem cell compendium. The first column represents the author’s name used as a reference for the study. The second column represents the stem cell type in which the experiment has been classified. The third column identifies the more specific stem cell population name. The fourth column gives the number of upregulated Entrez Gene genes in each stem cell population, while the final column gives the citation to the original paper.

I also collected PGLs that represented the genes upregulated in 49 differen-
<table>
<thead>
<tr>
<th>Author</th>
<th>Class</th>
<th>Primary stem cell</th>
<th># upregulated</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ivanova</td>
<td>Embryonic</td>
<td>ESC</td>
<td>666</td>
<td>[79]</td>
</tr>
<tr>
<td>RSantos</td>
<td>Embryonic</td>
<td>ESC</td>
<td>1365</td>
<td>[141]</td>
</tr>
<tr>
<td>Fortunel</td>
<td>Embryonic</td>
<td>ESC</td>
<td>1276</td>
<td>[56]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Embryonic</td>
<td>F-ESC</td>
<td>633</td>
<td>[1]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Embryonic</td>
<td>G-ESC</td>
<td>1784</td>
<td>[1]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Embryonic</td>
<td>N-ESC</td>
<td>1467</td>
<td>[1]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Embryonic</td>
<td>Z-ESC</td>
<td>1698</td>
<td>[1]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Embryonic</td>
<td>P-ESC</td>
<td>1109</td>
<td>[1]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Embryonic</td>
<td>iPSCbxo</td>
<td>1739</td>
<td>[1]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Embryonic</td>
<td>iPSC</td>
<td>1765</td>
<td>[1]</td>
</tr>
<tr>
<td>Hirst</td>
<td>Embryonic</td>
<td>ESC</td>
<td>500</td>
<td>[73]</td>
</tr>
<tr>
<td>Sharov</td>
<td>Embryonic</td>
<td>ES/EG</td>
<td>97</td>
<td>[156]</td>
</tr>
<tr>
<td>Sharova</td>
<td>Embryonic</td>
<td>ESC/EGC</td>
<td>1462</td>
<td>[157]</td>
</tr>
<tr>
<td>Tanaka</td>
<td>Embryonic</td>
<td>ES/TS</td>
<td>23</td>
<td>[173]</td>
</tr>
<tr>
<td>Ivanova</td>
<td>Neural</td>
<td>NSC</td>
<td>721</td>
<td>[79]</td>
</tr>
<tr>
<td>RSantos</td>
<td>Neural</td>
<td>NSC</td>
<td>1839</td>
<td>[141]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Neural</td>
<td>NSC</td>
<td>800</td>
<td>[1]</td>
</tr>
<tr>
<td>Fortunel</td>
<td>Neural</td>
<td>NSC</td>
<td>1315</td>
<td>[56]</td>
</tr>
<tr>
<td>Easterday</td>
<td>Neural</td>
<td>NS</td>
<td>65</td>
<td>[49]</td>
</tr>
<tr>
<td>Karsten</td>
<td>Neural</td>
<td>NSC</td>
<td>125</td>
<td>[88]</td>
</tr>
<tr>
<td>Buchstaller</td>
<td>Neural</td>
<td>NCSC</td>
<td>190</td>
<td>[26]</td>
</tr>
<tr>
<td>Aiba</td>
<td>Trophoblast</td>
<td>TSC</td>
<td>1208</td>
<td>[1]</td>
</tr>
<tr>
<td>Chateauvieux</td>
<td>Mesenchymal</td>
<td>MSC</td>
<td>254</td>
<td>[35]</td>
</tr>
<tr>
<td>Sharov</td>
<td>Mesenchymal</td>
<td>MS/NS</td>
<td>24</td>
<td>[156]</td>
</tr>
<tr>
<td>Ochsner</td>
<td>Liver</td>
<td>BMEL</td>
<td>874</td>
<td>[122]</td>
</tr>
<tr>
<td>Behbod</td>
<td>Mammary</td>
<td>MG-SP</td>
<td>269</td>
<td>[14]</td>
</tr>
<tr>
<td>Oatley</td>
<td>Spermatogonia</td>
<td>SSC</td>
<td>176</td>
<td>[121]</td>
</tr>
</tbody>
</table>

Table 5.2: List of cultured mouse stem cell profiling studies used in the mouse stem cell compendium. The first column represents the author’s name used as a reference for the study. The second column represents the stem cell type in which the experiment has been classified. The third column identifies the more specific stem cell population name. The fourth column gives the number of upregulated Entrez Gene genes in each stem cell population, while the final column gives the citation to the original paper.
tiated cell populations from the same studies and grouped the lists according to the stem cell from which the differentiated cell was derived. For example, all blood system populations were grouped together and labeled as expressed in cells derived from hematopoietic stem cells.

5.2.1.2 Establishment of replicate sets in mouse dataset compendium

One aspect that most previous expression-based stemness studies did not account for was the redundancy between individual studies. In assessing the contribution of independent experiments, this step is crucial as it ensures that very similar data sources are not counted twice. It is especially important to account for redundancy if the source of similarity is technical, rather than biological.

To account for the redundancy between studies, I initially evaluated the similarity between individual experiments by measuring the overlap of the genes upregulated in every pair of experiments. In more concrete terms, for every pair of experiments, starting from \((U_1, U_2)\) to \((U_{48}, U_{49})\), I evaluated the significance of overlap between the experiments in the pair as estimated using the hypergeometric distribution. To decide what experiments would be grouped together in a replicate set, if a pair \((U_{j1}, U_{j2})\) showed a similarity stronger than \(p = 10^{-50}\), it was kept for further examination.

I also tried less stringent p-value cutoffs \((p = 10^{-5}, p = 10^{-10}, \text{ and } p = 10^{-20})\), but they were not sufficiently stringent to separate experiments of the same stem cell type from experiments of different stem cell types. More stringent cutoffs separated not only the experiments of the different stem cell types, but also most of the experiments.
of the same stem cell type.

The p-value cutoff may look misleadingly stringent, but since the p-value is size-dependent and some experiments can show high level of overlap across cell types, it was a reasonable starting point. I used Cytoscape [155] to visualize the sets of connected components (experiments) and clustered the network using an edge-weighted spring-embedded layout, where the weight represented the strength of the p-value.

The final replicate sets (Figure 5.5) were selected based on their higher level of similarity within the same cell type and study, as compared to other cell types. A few experiments were borderline and could be associated with a replicate set (i.e., the Ramalho-Santos HSC signature), but since they represented experiments associated with a different study from the rest of the replicate set list, they were excluded from the final replicate set.

5.2.2 Human profiling studies

5.2.2.1 Study descriptions

In human, I collected published gene lists of significantly upregulated genes from 33 studies for 49 populations, corresponding to 9 general cell types. The first author names, publications, cell types and the numbers of upregulated genes in each study have been summarized in Tables 5.3 (cultured cell data) and 5.4 (primary cell data).

I also collected 38 PGLs of differentially downregulated genes in the same individual stem cell experiments. The downregulated PGLs represented 8 general stem
Figure 5.5: Gene-based similarities between all experiments in mouse stem cell compendium. Nodes represent individual experiments, while edges are drawn between nodes if the similarity between the two experiments exceed a p-value cutoff. Nodes are colored by stem cell type, cross-cell type edges are marked in red, while within-cell type links can be identified in blue. We observed many links at higher levels of similarity and selected replicate sets based on the closer similarity of the experiments to each other than to any other experiments. The final replicate sets are circled in black.

cell types.

5.2.2.2 Establishment of replicate sets in human dataset compendium

I assessed the similarity between individual human stem cell experiments using the exact same techniques and cutoffs as in the mouse stemness meta-analysis. Interestingly, the human data showed a clear separation between same and different cell types. In particular, while some mouse experiments showed a higher similarity to other experiments from the same lab and different cell type than to other experiments in the same cell type, the human experiments that showed a more significant overlap than $10^{-50}$
<table>
<thead>
<tr>
<th>Author</th>
<th>Class</th>
<th>Cultured stem cell</th>
<th># upregulated</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armstrong</td>
<td>Embryonic</td>
<td>ESC</td>
<td>156</td>
<td>[6]</td>
</tr>
<tr>
<td>Enver</td>
<td>Embryonic</td>
<td>Normal-ESC</td>
<td>2520</td>
<td>[52]</td>
</tr>
<tr>
<td>Enver</td>
<td>Embryonic</td>
<td>Adapted-ESC</td>
<td>1997</td>
<td>[52]</td>
</tr>
<tr>
<td>Player</td>
<td>Embryonic</td>
<td>ESC</td>
<td>847</td>
<td>[136]</td>
</tr>
<tr>
<td>Xu</td>
<td>Embryonic</td>
<td>ESC</td>
<td>1793</td>
<td>[198]</td>
</tr>
<tr>
<td>Calhoun</td>
<td>Embryonic</td>
<td>ESC</td>
<td>69</td>
<td>[29]</td>
</tr>
<tr>
<td>Cao</td>
<td>Embryonic</td>
<td>ESC</td>
<td>1489</td>
<td>[30]</td>
</tr>
<tr>
<td>Brandenberger</td>
<td>Embryonic</td>
<td>ESC</td>
<td>477</td>
<td>[25]</td>
</tr>
<tr>
<td>Brandenberger</td>
<td>Embryonic</td>
<td>ESC-MPSS</td>
<td>149</td>
<td>[24]</td>
</tr>
<tr>
<td>Cai</td>
<td>Embryonic</td>
<td>ESC</td>
<td>226</td>
<td>[28]</td>
</tr>
<tr>
<td>Sato</td>
<td>Embryonic</td>
<td>ESC</td>
<td>823</td>
<td>[152]</td>
</tr>
<tr>
<td>Skottman</td>
<td>Embryonic</td>
<td>ESC</td>
<td>163</td>
<td>[159]</td>
</tr>
<tr>
<td>Sperger</td>
<td>Embryonic</td>
<td>ESC</td>
<td>1174</td>
<td>[164]</td>
</tr>
<tr>
<td>Beqqali</td>
<td>Embryonic</td>
<td>ESC</td>
<td>98</td>
<td>[15]</td>
</tr>
<tr>
<td>Boquest</td>
<td>AdiposeStromal</td>
<td>ADSC</td>
<td>252</td>
<td>[19]</td>
</tr>
<tr>
<td>Donutu</td>
<td>Mammary</td>
<td>MamSC</td>
<td>77</td>
<td>[48]</td>
</tr>
<tr>
<td>Wright</td>
<td>Neural</td>
<td>NSC</td>
<td>370</td>
<td>[195]</td>
</tr>
<tr>
<td>Huang</td>
<td>Neural</td>
<td>NSC</td>
<td>1251</td>
<td>[74]</td>
</tr>
</tbody>
</table>

Table 5.3: List of cultured human stem cell profiling studies collected in human stem cell compendium. The first column represents the author’s name used as a reference for the study. The second column represents the stem cell type in which the experiment has been classified. The third column identifies the more specific stem cell population name. The fourth column gives the number of upregulated Entrez Gene genes in each stem cell population, while the final column gives the citation to the original paper.
<table>
<thead>
<tr>
<th>Author</th>
<th>Class</th>
<th>Primary stem cell</th>
<th># upregulated</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eckfeldt</td>
<td>Hematopoietic</td>
<td>BM-HSC</td>
<td>1623</td>
<td>50</td>
</tr>
<tr>
<td>Eckfeldt</td>
<td>Hematopoietic</td>
<td>UCB-HSC</td>
<td>1022</td>
<td>50</td>
</tr>
<tr>
<td>Jaatinen</td>
<td>Hematopoietic</td>
<td>HSC</td>
<td>317</td>
<td>[80]</td>
</tr>
<tr>
<td>Huang</td>
<td>Hematopoietic</td>
<td>CD133-HSC</td>
<td>840</td>
<td>[74]</td>
</tr>
<tr>
<td>Hemmoranta</td>
<td>Hematopoietic</td>
<td>CD133-HSC</td>
<td>152</td>
<td>[70]</td>
</tr>
<tr>
<td>Hemmoranta</td>
<td>Hematopoietic</td>
<td>CD34-HSC</td>
<td>111</td>
<td>[70]</td>
</tr>
<tr>
<td>Komor</td>
<td>Hematopoietic</td>
<td>HSC-Er</td>
<td>50</td>
<td>[97]</td>
</tr>
<tr>
<td>Komor</td>
<td>Hematopoietic</td>
<td>HSC-Gr</td>
<td>28</td>
<td>[97]</td>
</tr>
<tr>
<td>Kim</td>
<td>Hematopoietic</td>
<td>HSC-Megak</td>
<td>34</td>
<td>[97]</td>
</tr>
<tr>
<td>Kim</td>
<td>Hematopoietic</td>
<td>HSC-CD4T</td>
<td>287</td>
<td>[93]</td>
</tr>
<tr>
<td>Kim</td>
<td>Hematopoietic</td>
<td>HSC-CD8T</td>
<td>263</td>
<td>[93]</td>
</tr>
<tr>
<td>Kim</td>
<td>Hematopoietic</td>
<td>HSC-NK</td>
<td>181</td>
<td>[93]</td>
</tr>
<tr>
<td>Kim</td>
<td>Hematopoietic</td>
<td>HSC-Mye</td>
<td>482</td>
<td>[93]</td>
</tr>
<tr>
<td>Kim</td>
<td>Hematopoietic</td>
<td>HSC-Mono</td>
<td>334</td>
<td>[93]</td>
</tr>
<tr>
<td>Kim</td>
<td>Hematopoietic</td>
<td>HSC-ImmDendr</td>
<td>335</td>
<td>[93]</td>
</tr>
<tr>
<td>Kim</td>
<td>Hematopoietic</td>
<td>HSC-MatDendr</td>
<td>170</td>
<td>[93]</td>
</tr>
<tr>
<td>Wagner</td>
<td>Hematopoietic</td>
<td>HSC</td>
<td>23</td>
<td>[187]</td>
</tr>
<tr>
<td>Toren</td>
<td>Hematopoietic</td>
<td>HSC-Cord</td>
<td>514</td>
<td>[178]</td>
</tr>
<tr>
<td>Toren</td>
<td>Hematopoietic</td>
<td>HSC-Per</td>
<td>432</td>
<td>[178]</td>
</tr>
<tr>
<td>Ivanova</td>
<td>Hematopoietic</td>
<td>HSC</td>
<td>262</td>
<td>[79]</td>
</tr>
<tr>
<td>Igreja</td>
<td>Endothelial</td>
<td>EPC</td>
<td>230</td>
<td>[76]</td>
</tr>
<tr>
<td>Huang</td>
<td>Mesenchymal</td>
<td>MSC</td>
<td>604</td>
<td>[74]</td>
</tr>
<tr>
<td>Tsai</td>
<td>Mesenchymal</td>
<td>MSC</td>
<td>48</td>
<td>[179]</td>
</tr>
<tr>
<td>Kulterer</td>
<td>Mesenchymal</td>
<td>MSC</td>
<td>273</td>
<td>[103]</td>
</tr>
<tr>
<td>Song</td>
<td>Mesenchymal</td>
<td>MSC-AS</td>
<td>201</td>
<td>[162]</td>
</tr>
<tr>
<td>Song</td>
<td>Mesenchymal</td>
<td>MSC-OS</td>
<td>176</td>
<td>[162]</td>
</tr>
<tr>
<td>Song</td>
<td>Mesenchymal</td>
<td>MSC-CS</td>
<td>102</td>
<td>[162]</td>
</tr>
<tr>
<td>Kocer</td>
<td>Epithelial</td>
<td>EpSC</td>
<td>1091</td>
<td>[95]</td>
</tr>
<tr>
<td>Roh</td>
<td>Epithelial</td>
<td>EpSC</td>
<td>116</td>
<td>[147]</td>
</tr>
<tr>
<td>Kosinski</td>
<td>Intestinal</td>
<td>InEpSC</td>
<td>282</td>
<td>[100]</td>
</tr>
</tbody>
</table>

Table 5.4: List of non-cultured (primary) human stem cell profiling studies collected in human stem cell compendium. The first column represents the author's name used as a reference for the study. The second column represents the stem cell type in which the experiment has been classified. The third column identifies the more specific stem cell population name. The fourth column gives the number of upregulated Entrez Gene genes in each stem cell population, while the final column gives the citation to the original paper.
Figure 5.6: Gene-based global similarities between all experiments in human stem cell compendium. Nodes represent individual experiments, while edges are drawn between nodes if the similarity between the two experiments exceed a p-value cutoff. Nodes are colored by stem cell type and separate exactly into same cell-type cliques, which are used as the final replicate sets. The display uses an edge-weighted spring-embedded layout. “Replicate” sets are formed only for hematopoietic (red) and embryonic (light blue) stem cells.

However, this phenomenon was not necessarily biological in nature, but could be caused by technical reasons, such as similarities between the protocols and platforms used by the studies in each clique, especially since one of the cliques consisted of experiments from the same study. The cliques identified in the human data represented the final replicate sets I used in the human stemness meta-analysis.
5.3 Input gene modules

Gene modules represented the other main input to the stemness meta-analysis method, as gene-level based analyses had previously failed to identify shared stem cell mechanisms at the expression level. One of the primary hypotheses that I was interested in testing was whether functional redundancy and tissue-specific expression could explain previous failures to identify common stem cell mechanisms. Both functional redundancy and tissue-specific expression could arise through gene duplications, which led me to test putatively paralogous gene sets (labeled homolog modules hereafter) for significant stem cell expression.

Unfortunately, genome-wide predictions for mouse paralog gene sets were not readily available already, so I used a simple technique to identify such input gene modules.

5.3.1 Homolog gene families (modules)

I used BLASTP at an e-value\(^1\) cutoff of 0.05 to align the entire mouse proteome (mm9; 45,480 protein sequences\(^2\)) for each available mouse EntrezGene ID. BLASTP was an appropriate choice for this analysis, as I did not expect to need the recognition of very remote homologies.

For each pair of proteins, only the alignment with highest alignment score and e-value, as well as the highest overall sequence coverage was chosen as representative.

---

\(^1\)An e-value represents a value that measures the number of random hits we expect to see randomly in a given database.

\(^2\)Some proteins are represented by several different protein sequences, which may explain why the number is closer to 45,000, rather than to 25,000 sequences.
of the gene pair. After the initial screen, I only kept gene pairs whose sequences had an alignment e-value smaller than 10e-70 and coverage of more than 50%. Genes were connected to each other only if they satisfied this stringent requirement. I performed a depth-first traversal of the protein-protein similarity network to identify all connected components, each of which was thereafter used as a putative homolog family.

After the initial homolog family assignment, I performed a neighbor expansion step on all singleton gene families. Specifically, to incorporate more evolutionarily distant homology, unconnected genes were assigned to the family containing the gene with the most similar protein sequence if this best match exceeded a less stringent cutoff. I used an e-value cutoff of 10e-10 and a coverage cutoff of 50%. The process was iteratively repeated until no more singleton families could be reassigned to larger gene sets and the process converged on to a final set of homolog families (Figure 5.7). Genes remaining unconnected even after this expansion step were treated as singletons or singleton families in all subsequent analyses.

It should be noted that while this methodology works well for most homolog modules, it also has some pitfalls which can affect the homolog module quality. The use of depth-first traversal can cause “chains” of unrelated proteins to be placed in the same module. For example, protein A could be highly similar to protein B, and protein B could be highly similar to protein C without protein A sharing any similarity with protein C (Figure 5.8).

Summary results with a breakdown of homolog module number and sizes for each organism are provided in Tables 5.5 (for mouse) and 5.6 (for human).
Figure 5.7: Protein similarity network generation approach. BLASTP is used to generate alignments and at a stringent cutoff, depth-first search (DFS) is applied to identify all connected components: homolog families. Subsequently, an iterative neighbor expansion technique is applied to all singletons (red) until the set of homolog families converges to its final form.

Figure 5.8: Possible pitfalls of the homolog module definition methodology. The putative domain structure of four proteins (A, B, C, and D) is shown in the form of different shapes. The four proteins belong to the same homolog module. Each colored shape (triangle, square, diamond, circle) represents a different domain. While protein A is similar to protein B, and protein B is similar to protein C, protein A shares a low level of similarity with protein C. In addition, even though protein C and protein D share some similarity with protein A, they share no similarity with each other.
### Description of mouse module

<table>
<thead>
<tr>
<th>Description of mouse module</th>
<th>Before NE</th>
<th>After NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homolog groups with ≤ 100 members</td>
<td>3016</td>
<td>4653</td>
</tr>
<tr>
<td>Homolog groups with &gt;100 members</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Singleton genes</td>
<td>11920</td>
<td>5249</td>
</tr>
<tr>
<td><strong>Homolog gene families — Total</strong></td>
<td>14941</td>
<td><strong>9908</strong></td>
</tr>
</tbody>
</table>

Table 5.5: Summary of the distribution of mouse homolog (evolutionary) gene modules used as input to the stemness meta-analysis method. The neighbor expansion (NE) step significantly reduces the number of singleton genes represented in the homolog family input. The breakdown of all input module types can be found under the rightmost column. The total number of mouse homolog modules is shown in bold.

### Description of human module

<table>
<thead>
<tr>
<th>Description of human module</th>
<th>Before NE</th>
<th>After NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homolog groups with ≤ 100 members</td>
<td>2772</td>
<td>4346</td>
</tr>
<tr>
<td>Homolog groups with &gt;100 members</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Singleton genes</td>
<td>10983</td>
<td>4731</td>
</tr>
<tr>
<td><strong>Homolog gene families — Total</strong></td>
<td>13757</td>
<td><strong>9081</strong></td>
</tr>
</tbody>
</table>

Table 5.6: Summary of the distribution of human homolog (evolutionary) gene modules used as input to the stemness meta-analysis method. The total number of human homolog modules is shown in bold.

#### 5.3.2 Functional gene modules

While homolog modules could allow direct testing of the functional redundancy hypothesis, other functional relationships between genes could also be used to investigate common stem cell mechanisms. For example, protein complexes could be commonly used by most stem cell types. Alternatively, functional pathways, such as the Wnt or Notch signaling pathways could also be shared by multiple stem cell types. Therefore, I also evaluated the role of many different functional gene modules in the regulation of stem cell state.

I next summarize the methodology for the construction of the functional gene module set used in the mouse stemness study. The methodology in the human stemness...
study, however, is similar, so it will not be independently described.

I constructed a large set of functional modules, derived from five different data sources: GO [7], Biocarta, CORUM (experimentally derived mouse protein complexes) [148], and protein-protein interaction databases (modules identified from mouse protein-protein interaction data, as well as modules identified from human protein-protein interactions mapped to mouse genes by the corresponding best-reciprocal BLAST hit) [11, 134, 166, 197].

I compiled all of these modules together and excluded any gene sets that had more than 25% gene overlap with any other functional module. This filtering step aimed to reduce the level of redundancy inherent to many of these functional modules, because of

- the hierarchical nature of Gene Ontology,

- the high level of expected overlap between experimentally derived protein complexes and the predicted highly connected mouse and human protein-protein interaction modules.

Subsequently, I also removed any functional modules that showed more than 50% overlap with any homolog family (singletons excluded). The purpose of this filtering step was to allow the discovery of functional relationships between genes that could not have been identified using evolutionary relations. The relatively high overlap cutoff (50%) between evolutionary and functional gene modules was set to allow genes with multiple functional roles to be captured by different module types.
Table 5.7: Summary of the distribution of mouse functional gene modules used as input to the stemness meta-analysis method. The breakdown of all input module types can be found under the rightmost column. The total number of modules is shown in bold. Most mouse protein-protein interaction (PPI) modules have already been captured by experimentally derived protein complexes and removed for redundancy purposes, which explains the low number of these modules in the input data. The human PPI modules represent genes mapped to mouse gene space.

<table>
<thead>
<tr>
<th>Description of functional module</th>
<th># of modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimentally-derived protein complexes</td>
<td>90</td>
</tr>
<tr>
<td>Mouse PPI modules</td>
<td>5</td>
</tr>
<tr>
<td>Human PPI modules</td>
<td>140</td>
</tr>
<tr>
<td>GO, KEGG and Biocarta modules</td>
<td>376</td>
</tr>
<tr>
<td><strong>Functional gene modules — Total</strong></td>
<td><strong>611</strong></td>
</tr>
</tbody>
</table>

Table 5.8: Summary of the distribution of human functional gene modules used as input to the stemness meta-analysis method. The total number of modules is shown in bold.

<table>
<thead>
<tr>
<th>Description of functional module</th>
<th># of modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimentally-derived protein complexes</td>
<td>234</td>
</tr>
<tr>
<td>Human PPI modules</td>
<td>4</td>
</tr>
<tr>
<td>Mouse PPI modules</td>
<td>94</td>
</tr>
<tr>
<td>GO, KEGG and Biocarta modules</td>
<td>557</td>
</tr>
<tr>
<td><strong>Functional gene modules — Total</strong></td>
<td><strong>889</strong></td>
</tr>
</tbody>
</table>

The final set of modules from each source type is summarized in Tables 5.7 (for mouse) and 5.8 (for human).

In the next few sections I describe the steps of the SMA method in their application order – recurrence, diversity and specificity. Each section and scoring type will first begin with a notation section to allow the reader to follow the scoring descriptions more easily. The notations do not overlap, so each label will be uniquely associated with a single concept. Each notation section also includes a table for easy lookup of labels.
5.4 Recurrence scoring

The first step of the stemness meta-analysis method aims to measure the level of reproducibility and recurrence of upregulation associated with each gene module used as input (as described in the previous sections). Previous expression-based approaches to stemness have often taken simple approaches, such as vote-counting, to the integration of measurements across multiple experiments.

Simple approaches, however, are inappropriate if we need to integrate data from high-throughput gene expression experiments from multiple platform types. Platforms reached genome-wide coverage only relatively recently, so it is important to account for whether an individual gene was tested in a given experiment based on the input data.

Another important factor to consider is the signal strength of each experiment. While the use of differentially expressed genes is appropriate for cross-platform integration, simple integration approaches do not account explicitly for the differences in stringency and specificity associated with the identification of differentially expressed genes. Ideally, highly conservative experiments that identify a smaller fraction of upregulated genes will contribute a higher weight than experiments that reveal most tested genes to be significantly upregulated.

Furthermore, previous expression-based stemness approaches do not incorporate information about the similarities between experiments. As double-counting information from highly similar experiments is a concern (especially if the similarity is based
on technical reasons), it is crucial to account for the redundancy between individual experiments.

The recurrence score that I develop tries to address all of the above issues. To understand the score, however, I first give some notation definitions.

### 5.4.1 Notation definitions

Each module will be denoted with $M$, and the recurrence score for module $M$ will be labeled as $R(M)$. Module $M$ will consist of $n$ genes and each index over the genes in the module will be written as $i$. An expression compendium will consist of $k$ experiments, and each index over experiments will be indicated with $j$.

The fraction of upregulated genes in an experiment $j$ will be denoted by $v_j$, while $z_j$ will be used to indicate a “signal strength” weight for an experiment $j$.

To ease the understanding of the replicate sets, I will define a set of all replicate groups, $B_1, B_2, \ldots, B_F$, as a mutually exclusive and exhaustive partition of the experiments in the compendium, where $F$ is the number of replicate groups in the compendium. Each one of the replicate groups is generated as described in Section 5.2.1.2 and ranges in size between 1 and 7 experiments. Experiment $j$’s replicate group can be written as $\gamma(j)$, in which case if $\gamma(j) = 1$, then experiment $j \in B_1$. Finally, define $|B_{\gamma(j)}|$ to indicate the number of experiments in the “replicate” set $B_{\gamma(j)}$.

A summary of all recurrence score notation symbols presented here and in the following section can be found in Table 5.9.
<table>
<thead>
<tr>
<th>Notation symbol</th>
<th>Notation description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>Module of genes</td>
</tr>
<tr>
<td>$J$</td>
<td>Set of all experiments in expression compendium</td>
</tr>
<tr>
<td>$R(M)$</td>
<td>Recurrence score for module $M$</td>
</tr>
<tr>
<td>$n$</td>
<td>Number of genes in module $M$</td>
</tr>
<tr>
<td>$k$</td>
<td>Number of experiments in expression compendium</td>
</tr>
<tr>
<td>$F$</td>
<td>Number of replicate sets in expression compendium</td>
</tr>
<tr>
<td>$i$</td>
<td>Index over the genes in module $M$</td>
</tr>
<tr>
<td>$j$</td>
<td>Index over the experiments in the set $K$</td>
</tr>
<tr>
<td>$t_{ij}$</td>
<td>Tested/not-tested status of gene $i$ in experiment $j$</td>
</tr>
<tr>
<td>$x_{ij}$</td>
<td>Upregulation status of gene $i$ in experiment $j$</td>
</tr>
<tr>
<td>$v_j$</td>
<td>Fraction of upregulated genes in experiment $j$</td>
</tr>
<tr>
<td>$z_j$</td>
<td>“Signal strength” weight for experiment $j$</td>
</tr>
<tr>
<td>$B_{\gamma(j)}$</td>
<td>“Replicate” set for experiment $j$</td>
</tr>
<tr>
<td>$</td>
<td>B_{\gamma(j)}</td>
</tr>
<tr>
<td>$w_j$</td>
<td>“Replicate” weight for experiment $j$</td>
</tr>
<tr>
<td>$h_{i,\gamma(j)}$</td>
<td>Gene-experiment-specific “replicate” weight</td>
</tr>
</tbody>
</table>

Table 5.9: Overview of the recurrence score-associated notation. The first column represents the notation symbol used in the recurrence scoring definition. The second column gives the descriptions for all notation symbols introduced in the first column.

### 5.4.2 General form of a recurrence score

Let a module of genes be denoted as the set $M$. I define a recurrence score for a module, $R(M)$, to reflect the degree to which its constituent genes are upregulated in a collection of experiments, $J$. Incorporating all redundancy, signal strength, test status and upregulation status information, a gene-specific score for a gene $i$, $G_i$, can be written as follows:

$$G_i = \sum_{j \in J} z_j h_{i,\gamma(j)} x_{ij} t_{ij}.$$  \hspace{1cm} (5.1)

One can notice four elements that contribute to the overall score: $t_{ij}$ (whether a gene is tested or not), $x_{ij}$ (whether a gene is upregulated or not), $z_j$ (“signal strength” weight), and $h_{i,\gamma(j)}$ (“replicate” weight).
Both $t_{ij}$ and $x_{ij}$ are binary variables and can be described as simple measures of whether gene $i$ is respectively tested, or upregulated in experiment $j$:

$$t_{ij} = \begin{cases} 
1, & \text{if gene } i \text{ was tested in experiment } j; \\
0, & \text{otherwise.}
\end{cases}$$

$$x_{ij} = \begin{cases} 
1, & \text{if gene } i \text{ was upregulated in experiment } j; \\
0, & \text{otherwise.}
\end{cases}$$

$z_j$ represents the “signal strength” weight for an experiment $j$, such that more stringent experiments have a high signal strength, while less stringent experiments have a low signal strength. Various different measures can be used to determine the signal strength and I empirically determined a good choice for this weight, as shown in Section 5.4.3.

Finally, $h_{i,\gamma(j)}$ represents a gene-experiment-specific “replicate” weight. This weight is dependent on both the replicate set for experiment $j$ and the gene $i$, since different genes have been tested in different experiments, so the weight for a particular experiment is a function of the other experiments in which a gene was tested. The replicate weight can thus be expanded as follows:

$$h_{i,\gamma(j)} = \frac{w_j t_{ij}}{\sum_{l \in B_{\gamma(j)}} w_l t_{il}}, \quad (5.2)$$

where $w_j$ is the “replicate” weight associated just based on experiment $j$. 

87
After defining $G_i$, it is easy to also define a gene-specific normalization factor, $T_i$, that takes the replicate weight, the signal strength weight and the test status into account as

$$T_i = \sum_{j \in J} z_j h_{i,\gamma(j)} t_{ij}. \tag{5.3}$$

Given $T_i$ and $G_i$ for every gene in a module $M$, a recurrence score in its most general form can be written as follows:

$$R(M) = \sum_{i \in M} \frac{1}{T_i} (G_i)^q \tag{5.4}$$

$$= \sum_{i \in M} \frac{1}{\sum_{j \in J} z_j h_{i,\gamma(j)} t_{ij}} \left( \sum_{j \in J} z_j h_{i,\gamma(j)} x_{ij} t_{ij} \right)^q. \tag{5.5}$$

Here $q$ is the exponent associated with the gene score, determining how much contribution each gene will have to the final module score $R(M)$.

This general scoring scheme fits well with the meta-analysis techniques discussed in the previous chapters. It follows the vote-counting methods that count the number of experiments each gene is upregulated in, but it also incorporates score elements similar to the inverse variance weights typically used for combining effect sizes in standard meta-analysis.

### 5.4.3 Simulation of synthetic module data

Due to the lack of appropriate positive controls, I used synthetic data to assess the accuracy of each scoring method at identifying recurrent upregulation across many
stem cell types. Two distinct categories of synthetic modules were created: stemness and non-stemness modules. The data for each module type were synthesized from different underlying models.

5.4.3.1 Size of module selection

Initially, the size of each module was randomly sampled from an exponential distribution with a mean of four genes, which followed closely the distribution and average size of real homolog families. I generated data for 2000 modules, of which only 120 represented stemness families, based on my hypothesis that stemness families should represent no more than 5–6% of all tested families.

5.4.3.2 Module subtype selection

Three different stemness module subtypes, which reflected different possible patterns of upregulation, were simulated at equal ratios of 1:1:1, i.e. 40 stemness modules of each type. The three subtypes corresponded to the following patterns: families with predominant expression of a Single gene across Many tissues (type SM modules), families with expression of a Single member gene in a Single tissue (type SS modules), and families with a higher level of activity - Many genes expressed in Many tissues (type MM modules).

Two different non-stemness module subtypes—tissue-specific and unrelated—were simulated at a ratio of 2:3 (tissue-specific:unrelated). The tissue-specific subtype corresponded to families of genes that were primarily expressed in a single tissue, while
unrelated families represented families with no role in stem cell biology.

Data was simulated for nine experiments to match the experimental design of the founder studies, discussed previously in Section 4.1.1.

5.4.3.3 Sampling for each module subtype

Stemness module subtypes were sampled as follows:

- **Families with predominant expression of a single gene in many tissues (type SM modules).** Randomly select a single gene $i$ from the module and assign it a high probability of selection ($p_i=0.8$). Each other gene can be assigned the same low probability of selection estimated as \( \frac{0.2}{(n-1)} \), where $n$ is the total number of genes in the module. Since the type SM modules are expected to express predominantly a single gene, the probabilities of success for the genes in the module are heavily skewed towards a single gene. Then for each of the nine synthetic experiments, an independent single draw from a multinomial distribution is made with the above mentioned probabilities of success for each gene in the module.

- **Families with predominantly of a single gene in a single tissue (type SS modules).** Assign all genes in the module the same probability of selection, given as $\frac{1}{n}$, where $n$ is the number of genes in the module. Since the type SS modules are expected to express a single gene in a single tissue, the probabilities of success for each gene in the module are equal. Then for each of the nine synthetic
experiments, an independent single draw from a multinomial distribution is made, where all genes in the module have an equal probability of being picked.

- **Families with a higher level of activity - many genes expressed in many tissues (type MM modules).** Assign all genes in the module the same probability of selection, given as \( \frac{1}{n} \), where \( n \) is the number of genes in the module. Since the type MM modules are expected to express many gene in many tissues, the probabilities of success for each gene in the module are equal, but multiple draws are made in each tissue or experiment. Then for each of the nine synthetic experiments, two draws from the multinomial distribution are made, where all genes in the module have an equal probability of being picked.

Non-stemness module subtypes were sampled in a slightly different manner. First, for each non-stemness family, to decide which genes in the modules will be activated in any tissue type, a draw was initially made from a binomial distribution, \( \text{Binom}(n, p = 0.5) \), where \( n \) is the number of genes in the module. Once activated genes were selected, the sampling was performed as follows:

- **Tissue-specific families.** Randomly select a single tissue and assign it a high probability of selection (\( p_j = 0.8 \)). Each other tissue can be assigned the same low probability of selection estimated as \( \frac{0.2}{(k-1)} \), where \( k \) is the total number of experiments. Then, for each activated gene, make a single draw from a multinomial distribution with the above mentioned probabilities.

- **Families with no role in stem cells.** Assign all tissues the same probability
of selection, given as $\frac{1}{k}$, where $k$ is the total number of experiments. For each activated gene, make a single draw from a multinomial distribution, where all tissues have an equal probability of selection.

5.4.4 Evaluation and selection of a recurrence score

I used the simulated synthetic data to directly evaluate several recurrence scoring methods, each of which followed the general form of the recurrence score introduced in Equation 5.5 in Section 5.4.2.

I chose to vary two different parameters associated with the score. The first one was the exponent $q$ associated with the gene scores, which determined how much weight a gene will contribute to the final module score. I tested four different exponent choices: $q = 0.5, 1, 2$ and $3$. The second parameter was the “signal strength weight” $z_j$. I tested three different “signal strength” weight choices $z_j$: $-\log(v_j)$, $\frac{1}{v_j}$ and $(1 - v_j)$, where $v_j$ corresponded to the fraction of genes upregulated in experiment $j$.

I evaluated all combinations of the parameter choices, as well as an additional score that used a square as the exponent ($q = 2$), but did not use any form of experiment specificity weighting.

The results based on the simulated synthetic data indicated that the choice of signal strength weight was not as significant as the choice of the exponent associated with the gene scores. For all tested scoring methods within each exponent selection, no significant difference between the scores for different weights was observed, based on the Wilcoxon rank sum test.
To select an appropriate exponent, however, I calculated the area-under-the-curve (AUC) for different exponent choices, and compared the differences in AUC results, based on the stemness module type chosen for scoring (Table 5.10). The AUC represents a summary statistic usually derived from ROC (receiver-operating characteristic) plots, which compare the sensitivity and false positive rates associated with a scoring method. AUC values range between 0 and 1, and the best possible method will have the highest possible AUC.

The results indicated that high exponent choices, such as quadratic and cubic choices, scored highly with modules dominated by the expression of a single gene (type SM modules), while small exponent scoring choices, such as square root and linear selections scored the highest with modules showing a complex (type MM modules) or one-gene-one-tissue-like (type SS modules) stemness expression pattern.

One final AUC score was calculated as the average of the AUC scores for all stemness module types. The AUC scores for each individual stemness module type, as well as a summary of the average AUC score results are shown in Table 5.10.

I also directly compared the ranks of the scores of all simulated stemness modules between the different scoring types using the Wilcoxon rank sum test. The comparisons indicated no significant difference between the ranks of the scores, given by the quadratic and cubic scoring choices (Wilcoxon rank sum test: p-value = 0.8172). However, these scoring choices were both significantly different from the square root scoring (quadratic: p-value = 0.0048 and cubic: p-value = 0.00863).

Finally, I selected the experiment-weighted quadratic \(q = 2; z_j = -log(v_j)\)
Table 5.10: Area-under-curve (AUC) results for 13 recurrence scoring methods tested for recovery of stemness modules, based on synthetic data evaluation. Two scoring parameters were varied and each scoring method represented in this table shows a different pair of parameter choices. The row that describes the final parameter choice associated with the highest average AUC is shown in bold. The first column (Scoring method) summarizes the parameter choice for each score. The second column (Type SM AUC) gives the AUC results for stemness families with predominant expression of a single gene in many tissues (abbreviated as type SM). The third column (Type SS AUC) gives the AUC results for stemness families with predominant expression of a single gene in a single tissue (abbreviated as type SS). The fourth column (Type MM AUC) gives the AUC results for stemness families with a higher level of activity - multiple genes in multiple tissues (abbreviated as type MM). The fifth column shows the average AUC across the type SM, type SS, and type MM families. The highest AUCs in each type are shown in italics in each column. The critical parameter in the simulation is $q$. Smaller values of $q$ produce more accurate results for modules that use multiple genes in multiple tissues, while higher values of $q$ produce more accurate results for modules that use a single gene in many tissues.
recurrent scoring method as the method with the highest average area-under-the-curve among the tested scoring measures and used it in all subsequent analyses. The final form of the recurrent scoring can be summarized from Equation 5.5 as:

$$ R(M) = \sum_{i \in M} \sum_{j \in J} \frac{1}{-\log(v_j) h_{i,\gamma(j)} t_{ij}} \left( \sum_{j \in J} -\log(v_j) h_{i,\gamma(j)} x_{ij} t_{ij} \right)^2, \quad (5.6) $$

where all score elements remain as previously described.

5.4.5 Significance of recurrence scores

Once I selected a recurrence scoring method and applied it the real stem cell compendium data, I had to also perform an evaluation of the significance of recurrence scores. The significance of the recurrence scores was assessed using permutation analysis and was used to identify modules with significantly recurrent upregulation scores across different stem cell experiments.

The data within each experiment were permuted 1,000 times. The total number of upregulation counts in each experiment was kept the same, but the upregulation states were randomly permuted between genes and each gene acquired a new upregulation state assignment. This procedure was followed to preserve the distribution of the number of upregulated genes the same within each experiment and yet break the potential dependency in the expression of the genes associated with each module.

If an experiment was part of a replicate set, the block of data values for each gene within the replicate set was treated as one experiment and permuted together. I
inferred 1,000 randomized recurrence scores for each module using the score described in Equation 5.6. All module scores were normalized by the size of the module. Since the recurrence score was dependent on the size of the module and to avoid empirical distribution fits, I used an FDR approach to select an appropriate significance cutoff for each module size.

At the most general level, FDR can be defined as

\[
FDR = \frac{FP}{TP + FP}
\]  

(5.7)

While I didn’t know the number of false positives in this context, I could approximate their number from the randomly permuted data.

\[
FP = \# \text{ of modules } M \text{ in the randomly permuted data with } R(M) > c
\]  

(5.8)

\[
FP + TP = \# \text{ of modules } M \text{ in the real data with } R(M) > c
\]  

(5.9)

Therefore, the FDR could be computed from Equations 5.8 and 5.9 as

\[
FDR = \frac{\# \text{ of modules } M \text{ in the randomly permuted data with } R(M) > c}{\# \text{ of randomly permuted modules}}
\]  

\[
\times \frac{\# \text{ of modules } M \text{ in the real data with } R(M) > c}{\# \text{ of real modules}}
\]  

(5.10)

and used to select the cutoff \( c \) separately for each different module size. To ensure conservativeness, I selected a FDR cutoff of 5% and estimated the corresponding recurrence score cutoff for each family size (Table 5.11); the cutoff was thereafter used as a significance cutoff for that family size. The overall number of recurrent modules identified in each organism is shown in Figure 6.2 (in Chapter 6 for mouse) and Figure 7.2 (in Chapter 7 for human).
Table 5.11: Recurrence score cutoffs for modules of different sizes. The recurrence score cutoffs represent the recurrence scores for each module size at which the false discovery rate was 5%. Any modules with recurrence scores higher than the ones indicated in this table are considered significantly recurrent. The first column shows different module sizes. If the column has a single value, only one module size is shown in the row. If a range of values is introduced in the first column, all modules of these sizes have similar cutoffs. The second column shows the recurrence score cutoffs for the different module sizes.

<table>
<thead>
<tr>
<th>Module Size</th>
<th>Recurrence cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1.4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6-8</td>
<td>0.8</td>
</tr>
<tr>
<td>9-10</td>
<td>0.7</td>
</tr>
<tr>
<td>11</td>
<td>0.6</td>
</tr>
<tr>
<td>12</td>
<td>0.7</td>
</tr>
<tr>
<td>13-19</td>
<td>0.6</td>
</tr>
<tr>
<td>20-25</td>
<td>0.5</td>
</tr>
<tr>
<td>&gt; 25</td>
<td>0.1-0.3</td>
</tr>
</tbody>
</table>
5.5 Diversity scoring

Even though recurrence scoring could successfully identify families with recurrently upregulated member genes across a large set of stem cell experiments, the over-representation of certain stem cell types in the literature, such as embryonic stem cells, could lead to the discovery of recurrently upregulated families in only a single or a few stem cell types. To distinguish between modules with high and low level of stem cell type coverage, I followed a different, information-theoretic entropy-based scoring strategy.

Based on the input published gene list (PGL) differential expression data, every module has an associated binary expression matrix that represents the upregulation status of the individual member genes in every experiment. These upregulation matrices can be transformed into cell type-based matrices, such that each row represents an individual gene and each column represents an individual cell type. Every entry in the cell-type based matrix corresponds to the fraction of experiments in the cell type that the gene was upregulated in.

5.5.1 Notation definitions

Each module will be denoted with $M$, a cell type diversity score for module $M$ will be labeled as $d(M)$, and a gene usage diversity score will be labeled as $g(M)$.

The total number of tested cell types from the $k$ experiments used in the stem cell expression compendium will be denoted as $s$, and as previously described $n$ will be
<table>
<thead>
<tr>
<th>Notation symbol</th>
<th>Notation description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>Module of genes</td>
</tr>
<tr>
<td>$L$</td>
<td>Set of all stem cell types in compendium</td>
</tr>
<tr>
<td>$d(M)$</td>
<td>Cell type diversity score for module $M$</td>
</tr>
<tr>
<td>$g(M)$</td>
<td>Gene usage diversity score for module $M$</td>
</tr>
<tr>
<td>$n$</td>
<td>Number of genes in module $M$</td>
</tr>
<tr>
<td>$k$</td>
<td>Number of experiments in expression compendium</td>
</tr>
<tr>
<td>$s$</td>
<td>Number of stem cell types in expression compendium</td>
</tr>
<tr>
<td>$i$</td>
<td>Index over the genes in module $M$</td>
</tr>
<tr>
<td>$l$</td>
<td>Index over the stem cell types in expression compendium</td>
</tr>
</tbody>
</table>

Table 5.12: Overview of the diversity-score-associated notation. The first column represents the notation symbol used in the diversity scoring definition. The second column gives the descriptions for all notation symbols introduced in the first column.

used to reflect the total number of genes in module $M$. Each index over the genes in the module can still be written as $i$, and each index over cell types can be denoted with $l$.

The set of all tested cell types will be denoted with $L$.

A summary of all diversity-related notation symbols presented here can be found in Table 5.12.

### 5.5.2 Cell type and gene usage diversity

To understand why an entropy-based score is an appropriate choice for the measurement of diversity, I first introduce the concept of entropy. Information-theoretic entropy measures the level of uncertainty attached to a random variable. Let’s assume I would like to measure the entropy associated with some future event, based on information about the probabilities of the various event outcomes. If all outcomes in the outcome space are all equally probable, the uncertainty of the outcome of the future event is very high and the entropy score is the highest possible. Conversely, if the prob-
abilities in the outcome space are so skewed that only one of the outcomes is possible, the uncertainty of the outcome of the future event is minimal and the entropy score is the lowest possible. For a random variable A with \( m \) possible outcomes \( (a_1, a_2, ..., a_m) \), the entropy can be written as

\[
H(A) = - \sum_{i=1}^{m} p(a_i) \log(p(a_i)) . \tag{5.11}
\]

I estimated two different entropy-based diversity scores. The first one reflects the cell type diversity of the module, while the other reflects the gene usage diversity. For a family \( M \), I defined the cell type diversity (entropy) score \( d(M) \) as follows:

\[
d(M) = - \sum_{l \in L} \alpha_l(M) \log(\alpha_l(M)) , \tag{5.12}
\]

where \( \alpha_l(M) \) represents the fraction of upregulated genes from module \( M \) in cell type \( l \) and can be defined as

\[
\alpha_l(M) = \frac{\sum_{i \in M} f_{il}}{\sum_{l' \in L} \sum_{i \in M} f_{il'}} , \tag{5.13}
\]

and if \( Z(l) \) represents the set of all experiments of cell type \( l \), then \( f_{il} \) can be written as

\[
f_{il} = \frac{\sum_{j \in Z(l)} x_{ij}}{\sum_{j \in Z(l)} t_{ij}} . \tag{5.14}
\]

Here the binary variables \( x_{ij} \) and \( t_{ij} \) still represent the upregulation and test status respectively of a gene \( i \) in experiment \( j \).

Similarly, I defined the gene usage diversity (entropy) score \( g(M) \) as follows:
where $\beta_i$ corresponds to the fraction of cell types in which a gene $i$ is upregulated and can be defined as

$$\beta_i = \frac{\sum_{l \in M} f_{il}}{\sum_{l' \in M} \sum_{l \in L} f_{l'}}.$$  \hspace{1cm} (5.16)

$L$ represents the set of all stem cell types in the compendium, while $f_{il}$ is still used to describe the fraction of experiments of cell type $l$ in which a gene $i$ is upregulated.

I also estimated normalized cell type ($d_{\text{norm}}(M)$) and gene usage ($g_{\text{norm}}(M)$) diversity scores to take the maximum possible diversity (all possible outcomes are equiprobable) into consideration, as described by the denominators of both scores:

$$d_{\text{norm}}(M) = \frac{d(M)}{\log(s)}$$  \hspace{1cm} (5.17)

$$g_{\text{norm}}(M) = \frac{g(M)}{\log(n)}$$  \hspace{1cm} (5.18)

This normalization step is especially important for the gene usage diversity score, because it accounts for the differences between the sizes of different modules.

5.5.3 Significance of diversity scores

The diversity score significance assessment was performed in a similar fashion to the evaluation of the recurrence scores. Randomized families from 1,000 permutations were used to generate cell-type diversity scores for random modules. For each family
size, a false discovery rate was estimated at every cutoff from 0 to the maximum possible entropy score.

However, since the total number of tested cell types was a constant across modules of all sizes, and since empirical observations suggested that the FDR cutoffs for modules of different sizes would be relatively similar (see Figure 6.6), I generated a combined FDR estimate across all module sizes at each cutoff.

The combined estimate used a weighted average of the FDR levels associated with each size weighted by the number of families of that size. The final cell-type diversity score cutoff corresponded to the 5% FDR level cutoff; any family above it was considered cell type diverse. This cutoff was 2.5, where the maximum possible score that could be achieved for twelve stem cell types was 3.6.

Since gene diversity was not as crucial to the determination of stemness, but rather to the overall participation and contribution of individual member genes of the module, a simple cutoff was used to distinguish gene diverse modules from modules that exhibited low gene diversity. Families that showed a normalized gene diversity score higher than 0.5 were considered “gene-diverse” modules.

5.6 Specificity scoring

The discovery of modules that are reproducibly upregulated across many different stem cell types does not guarantee their specificity to undifferentiated cells. To identify and exclude any families that were also significantly upregulated across differ-
entiated cells, I defined a *specificity score*. The specificity score I chose was based on the recurrence score of the module across all differentiated cell experiments. Significance of the recurrence score of a module across the differentiated cells was determined using the 5% FDR cutoffs for modules of that size as identified from the stem cell data. If a module was significantly upregulated in the stem cell experiments, but not significantly upregulated in the differentiated cell experiments, it was considered “specific.” If a module was significantly recurrently upregulated in both stem and differentiated cells, it was considered non-specific and excluded from further consideration in both the stemness “on” and “off” categories.

### 5.7 Pattern classification

The three scoring elements described in the previous sections allow the classification of modules into several different classes. Because of the noise inherent to microarray data and the worry about differences between the experimental designs of individual experiments, only significant recurrently upregulated modules were considered for further classification.

Some interesting biological phenomena that are not reproducible across all conditions and cell populations could be missed because of this restriction, but I chose the more conservative statistically sound approach, trying to reduce the biological false positives at the expense of potential false negatives.

The most important distinction, as related to stemness, between individual
Figure 5.9: A pattern classification procedure used to identify stemness modules. The path to the definition of both types of stemness modules – all-for-all (AFA) and one-for-all (OFA) – is outlined in red.

The cell diversity score delineates two classes of modules: modules with significantly high cell diversity scores and modules with low cell diversity scores. Highly (and significantly) cell-type diverse modules can fall in one of four categories:

- **All-for-all (AFA) modules** upregulate many member genes in most stem cell types. They show significant specificity to stem cells and have a normalized gene diversity score higher than 0.5. These gene modules can vary from protein complexes, which require all member genes to be differentially upregulated in stem cells to homolog modules that exhibit multi-cell type specialization of individual
each gene can be highly upregulated in several stem cell types. AFA modules represent one of the two categories of gene sets that is associated with stemness.

- **One-for-all (OFA) modules** upregulate predominantly one gene in most stem cell types. They also show a significant specificity to stem cells, but they have a normalized gene diversity lower than 0.5. Even though many of these modules consist of more than one gene, there is only one predominant gene consistently upregulated in different stem cells. OFA modules represent the second category of gene sets that is associated with stemness.

- **Constitutive module (CM) sets** upregulate many member genes in most stem cell types and a significant set of differentiated cells as well. They are not specific to stem cells, but they have a normalized gene diversity score higher than 0.5. These modules consist of genes that are potentially required either for housekeeping roles in the cell, or alternatively have cell-type specific roles that are unrelated to the functional properties of stem cells.

- **Constitutive gene (CG) sets** upregulate predominantly one gene in most stem cell types and a significant number of differentiated cells. They are not specific to stem cells and have a normalized gene diversity score lower than 0.5. These modules are expected to be rare and theoretically consist of one predominant expressed gene in stem cells, and one or a combination of other genes expressed in differentiated cells.
Based on the definitions above and for completeness purposes we can also define several categories of non cell-type diverse modules. However, in practice most of these categories do not bear a clear relevance to stemness, so I restricted the pattern classification to only two non-cell type diverse module categories:

- **All-for-one (AFO) modules** upregulate many member genes in a low number of stem cells and can be described as cell-type-specific modules. They have normalized gene diversity higher than 0.5 and can show specificity to stem cells, but are not required to as the definition of non-specificity does not have a clear meaning at this level.

- **One-for-one (OFO) modules** upregulate few member genes in a low number of stem cells and can be described as single-gene-dominated cell-type-specific modules. They have normalized gene diversity lower than 0.5.

### 5.8 Stemness “on” and stemness “off” modules

I used the pattern classification procedures and the combination of all scores to define modules associated with stemness. Modules that fell in either the all-for-all (AFA), or one-for-all (OFA) categories were defined as stemness modules. As previously mentioned, while modules of upregulated genes in many stem cell types (stemness “on” modules) represented clear examples of gene sets with stem-cell-related functions, modules of downregulated genes across many stem cell types (stemness “off” or differentiation modules) could also be very important.
These stemness “off” modules could be identified in the exact same way as the stemness “on” modules, but using the differentiated cell data, instead of the stem cell data. Stemness “off” modules would then be defined as significantly recurrent across differentiated cell experiments, significantly cell-diverse across many differentiated cell types, and significantly specific to differentiated cells.

5.9 Formulation of stemness index

Stemness module identification is important for discovering families, modules and processes associated with self-renewal and the control of differentiation. Individual gene families and their potential regulatory roles in the maintenance of the stem cell state may provide testable hypotheses in existing stem cell types, as well as suggest potential markers for less well-studied stem cells. However, the SMA method can also be viewed as a feature selection method, which identifies the most relevant features (stemness families) that could be used to distinguish between stem cell and non-stem cell types.

In its most simple form, classification can be done through the calculation of a single stemness index score that measures how stem-cell-like the population is. A general advantage of such a score is that it avoids the explicit binary classification: stem cell/non-stem cell, as there is a range of cells at various stages of differentiation in between these two classes.
5.9.1 Notation definitions

I will denote the set of all stemness (“on”) modules with $S$, and the set of all differentiation (stemness “off”) modules with $D$. An individual module will still be denoted with $M$, and the total number of genes in module $M$ will still be denoted with $n$.

Each index over stemness modules can be written as $a$, while each index over the differentiation modules can be written as $b$. The total number of stemness modules will be labeled with $r$ and the total number of differentiation modules will be denoted with $u$.

The stemness index score will be written as $SI$, and each new experiment gene signature will be denoted with $E$. The activation of an individual module $M$ will be labeled as $A$, while the fraction of upregulated genes in a module will be denoted with $f$.

A summary of all stemness index-related notation symbols can be found in Table 5.13.

5.9.2 Cross-validation setup

To select the most accurate stemness index score, I used a cross-validation framework and specifically a 5-fold cross-validation setup based entirely on the mouse stem cell compendium data.

The basic cross-validation setup is shown in Figure 5.10.

The 49 stem cell populations were randomly assigned to five different groups,
Figure 5.10: Visual illustration of the five-fold cross-validation setup. The stem cell compendium is divided into five parts and each part uses 4/5 of the original data to identify stemness modules (black solid lines) and 1/5 of the original data (black dashed lines) to test and evaluate the stemness index scores.
where four groups consisted of ten experiments and one consisted of nine experiments. I then created five datasets from these groups, where in each dataset four of the groups were all used together as input to the SMA method, while the last group was used for testing. Thus, each dataset consisted of 39 or 40 input stem cell populations and the stemness score definition used only homolog modules as input. For each stem cell dataset I also defined the corresponding differentiated cell dataset. The differentiated cell dataset consisted of all differentiated cell experiments, associated with the particular stem cell experiments in that cross-validation subset.

Following the SMA method application described earlier, I identified stemness modules from each cross-validation set and the summary of the results is presented in Tables 5.14 and 5.15. Because of the random assignment of experiments to each of the initial five groups, some of the groups had a smaller number of actual stem cell types than the original twelve stem cell types. To define the stemness modules, I used the

<table>
<thead>
<tr>
<th>Notation symbol</th>
<th>Notation description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
<td>New experiment gene signature</td>
</tr>
<tr>
<td>$M$</td>
<td>Module of genes</td>
</tr>
<tr>
<td>$S$</td>
<td>Set of all stemness “on” modules</td>
</tr>
<tr>
<td>$D$</td>
<td>Set of all stemness “off” (differentiation) modules</td>
</tr>
<tr>
<td>$n$</td>
<td>Number of genes in module $M$</td>
</tr>
<tr>
<td>$r$</td>
<td>Number of stemness modules in $S$</td>
</tr>
<tr>
<td>$u$</td>
<td>Number of differentiation modules in $D$</td>
</tr>
<tr>
<td>$a$</td>
<td>Index over the stemness modules in $S$</td>
</tr>
<tr>
<td>$b$</td>
<td>Index over the differentiation modules in $D$</td>
</tr>
<tr>
<td>$SI$</td>
<td>Stemness index score</td>
</tr>
<tr>
<td>$A$</td>
<td>Activation score for a gene module</td>
</tr>
<tr>
<td>$f$</td>
<td>Fraction of upregulated genes in a module</td>
</tr>
</tbody>
</table>

Table 5.13: Overview of the stemness index score notation. The first column represents the notation symbol used in the stemness index scoring definition. The second column gives the descriptions for all notation symbols introduced in the first column.
<table>
<thead>
<tr>
<th>CV fold</th>
<th># recurrent stem cell homolog modules</th>
<th># stemness “on” homolog modules</th>
<th># recurrent stem cell functional modules</th>
<th># stemness “on” functional modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV fold 1</td>
<td>256</td>
<td>74</td>
<td>90</td>
<td>39</td>
</tr>
<tr>
<td>CV fold 2</td>
<td>225</td>
<td>82</td>
<td>81</td>
<td>53</td>
</tr>
<tr>
<td>CV fold 3</td>
<td>163</td>
<td>84</td>
<td>58</td>
<td>39</td>
</tr>
<tr>
<td>CV fold 4</td>
<td>163</td>
<td>50</td>
<td>65</td>
<td>43</td>
</tr>
<tr>
<td>CV fold 5</td>
<td>194</td>
<td>40</td>
<td>72</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 5.14: Stemness “on” modules identified in each cross-validation (CV) fold. Every row represents summary data for a different cross-validation subset. The first column defines the cross-validation folds. The second column (# recurrent stem cell homolog) shows the number of significantly recurrent homolog modules in the stem cell data from that cross-validation subset. The third column (# stemness “on” homolog modules) shows the number of stemness “on” homolog modules inferred from the stem cell data. The fourth column (# recurrent stem cell homolog) shows the number of significantly recurrent functional modules in the stem cell data from that cross-validation subset. The fifth column (# stemness “on” functional modules) shows the number of stemness “on” functional modules identified from the stem cell data.

previously established significance cutoffs for recurrence, but adjusted the significance cutoffs for the cell-type diversity to proportionally match the original cutoff based on the number of cell types used in training. For example, if the original cell diversity cutoff was $d = 2.5$ and the new subset of experiments consisted of only ten stem cell types instead of twelve, the cutoff was adjusted to $d = 2.5 \times \frac{\log(10)}{\log(12)}$.

To ensure that the stemness and differentiation modules inferred from each cross-validation (CV) fold were robust to the experiment selection, I compared the recurrence scores for each module across all cross-validation sets. If the modules were highly robust, they would show similar recurrence scores across all cross-validation sets. Indeed, this was the result I observed. Most modules showed highly similar scores across most cross-validation sets (Figure 5.11), suggesting that the stemness and differentiation modules identified from each cross-validation set are unlikely to be highly affected by
Table 5.15: Differentiation families identified in cross-validation (CV) fold. Every row represents summary data for a different cross-validation subset. The first column defines the cross-validation sets. The second column (# recurrent diff. cell homolog) shows the number of significantly recurrent homolog modules in the differentiated cell data from that cross-validation subset. The third column (# stemness “off” homolog modules) shows the number of stemness “off” (differentiation) homolog modules inferred from the differentiated cell data. The fourth column (# recurrent diff. cell functional) shows the number of significantly recurrent functional modules in the differentiated cell data from that cross-validation subset. The fifth column (# stemness “off” functional modules) shows the number of stemness “off” (differentiation) functional modules identified from the differentiated cell data.

<table>
<thead>
<tr>
<th>CV fold</th>
<th># recurrent diff. cell homolog</th>
<th># stemness “off” homolog modules</th>
<th># recurrent diff. cell functional</th>
<th># stemness “off” functional modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV fold 1</td>
<td>210</td>
<td>44</td>
<td>42</td>
<td>16</td>
</tr>
<tr>
<td>CV fold 2</td>
<td>142</td>
<td>27</td>
<td>27</td>
<td>11</td>
</tr>
<tr>
<td>CV fold 3</td>
<td>93</td>
<td>24</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>CV fold 4</td>
<td>81</td>
<td>8</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>CV fold 5</td>
<td>218</td>
<td>16</td>
<td>39</td>
<td>6</td>
</tr>
</tbody>
</table>

the experiment selection in each cross-validation set.

5.9.3 Binary switch and fractional scoring

One naive approach for measuring the association of a new experiment with stemness is to summarize each stemness module by a single value, which captures the activation of the module in a given experiment. Two primary candidates include a “binary switch” score and a “fraction of gene activation” score. Given a new set of differentially expressed genes, the binary switch score considers a module activated if there is at least one gene in the module that has been identified as differentially expressed. The fractional score, on the other hand, estimates the activation of a module as the fraction of the genes in that module that have been marked as differentially
Figure 5.11: Robustness of recurrence scores across cross-validation (CV) folds. The heatmap shows the recurrence scores of all mouse homolog modules in each cross-validation fold. Each column represents a different cross-validation fold. Each row represents a different homolog module. To ease viewing, only modules that had a non-zero recurrence score in at least one cross-validation fold were plotted. High recurrence scores are shown in yellow, while the low (or zero) recurrence scores are shown in black.
expressed.

To combine the activation input from all stemness modules, we can estimate the average stemness module activation score. Alternatively, we can define a weighted average stemness module activation score, where the weight contribution associated with each stemness module is the cell-diversity of the module, based on the results of the application of the original SMA method.

These possibilities suggest four different scoring combinations: a weighted fractional, weighted binary, unweighted fractional, and an unweighted binary score. The formulas for the stemness index $SI$ of a new experiment gene signature $E$ is described in Equation 5.19 (unweighted scoring combinations) and Equation 5.20 (weighted scoring combinations):

$$SI(E) = \frac{\sum_{a \in S} A_a(E)}{r}$$  \hspace{1cm} (5.19)

$$SI(E) = \frac{\sum_{a \in S} d(a) A_a(E)}{\sum_{a \in S} d(a)}$$  \hspace{1cm} (5.20)

$$A_a(E) = \begin{cases} 
1, & \text{if binary score;} \\
 f_a(E), & \text{if fractional score},
\end{cases}$$

where $d(a)$ is the cell diversity score associated with each stemness module $a$, and $f_a(E)$ is the fraction of genes in module $a$ that are upregulated and are represented in the experiment gene signature $E$ (i.e., number of genes in module $a$ that are upregulated,
divided by the total number of genes in the module).

All four stemness index scores – binary unweighted, binary weighted, fractional unweighted and fractional weighted were evaluated on each of the cross-validation sets. Based on the stem cell (SC) experiments and differentiated cell (DC) experiments and only the homolog-based stemness features, I defined the precision and recall as follows:

\[
\text{Precision} = \frac{\#\text{SC exp. with index} > c}{\left( \#\text{SC exp. with index} > c \right) + \left( \#\text{DC exp. with index} > c \right)} \tag{5.21}
\]

\[
\text{Recall} = \frac{\#\text{SC exp. with index} > c}{\left( \#\text{SC exp. with index} > c \right) + \left( \#\text{SC exp. with index} \leq c \right)} \tag{5.22}
\]

I swept through the cutoffs \(c\) and compared the precision-recall curves associated with each score. Initial observations suggested high similarity between all four scores, though the fractional weighted and unweighted scoring measures were most similar and performed slightly better than their binary counterparts (Figure 5.12).

I also evaluated the precision of each stemness index score directly as a function of the score (Figure 5.13) and observed the highest correlation between the precision and index score (\(\rho = 0.869\)) using the weighted fractional score.

While these stemness index scoring measures captured some stem-cell-associated signal, overall their performance was relatively poor. This poor performance could be due to the lack of sufficient signal in the activation scores, which were the most significantly contributing source of information to the stemness index scores. In addition, consideration of not only stemness, but differentiation modules as well could have improved the overall accuracy of these measures. In this context, I examine a more sophisticated
Figure 5.12: Precision-recall comparison of four stemness index scores - a binary un-weighted (green dashed line), a binary weighted (blue dashed line), a fractional un-weighted (red dashed line), and a fractional weighted (black solid line) scoring measures. The calculation of the stemness indices is based on stemness “on” homolog features only. Recall is plotted on the x-axis, while precision is shown on the y-axis. The most accurate method is the one approximately closest to the upper right-hand-side corner.
Figure 5.13: Precision comparison of four stemness index scores. The upper panel shows the precision comparison between the binary unweighted (green dashed line) and a binary weighted (blue solid line) scoring measures. The lower panel shows the precision comparison between the fractional unweighted (red dashed line), and a fractional weighted (black solid line) scoring measures. Each x-axis shows directly the values produced by the scoring measures plotted in that panel. The y-axis shows the precision value.
stemness index approach next.
5.9.4 Log-likelihood approach

Based on the methodology described earlier, the SMA method can identify not only stemness modules, but also differentiation modules. A more sophisticated approach would incorporate not only the stemness modules as features, but the differentiation modules as well. Such a score would be naturally balanced around zero, where positive scoring experiments are more stem cell-like, while negative scoring experiments are more differentiated cell-like.

It is also natural to incorporate all identified features, homolog and functional, and compare the classification accuracy associated with the use of each type.

Each module can still be weighted by its own cell diversity, while the activation element of the score can be associated with a value that captures the level of enrichment of the module in the new experiment. While the fraction of upregulated genes described as a possible choice in Section 5.9.3 carries some information about the enrichment of the module, we also want to account for the module size and the deviation from expectation. I consider two other related module activation scores:

- the log-ratio of the observed fraction of activated genes to the expected fraction of activated genes, as given by the hypergeometric distribution (log-ratio method)

- the significance of overlap between the module and the genes upregulated in the new experiment, as measured by a -log10 p-value score (p-value method)

I incorporated all of these considerations into the formulation of four different scores, mathematically formulated as shown in Equations 5.23–5.26. The first two equa-
tions (Equations 5.23–5.24) represent the log-ratio method, while Equations 5.25–5.26 correspond to the p-value method. All four formulas, however, build on the weighted fractional scoring introduced earlier in Equation 5.20.

\[
SI(E) = \frac{\sum_{a \in S} d(a) \log(\frac{\text{Observed}[f_a(E)]}{\text{Expected}[f_a(E)]})}{\sum_{a \in S} d(a)} - \frac{\sum_{b \in D} d(b) \log(\frac{\text{Observed}[f_b(E)]}{\text{Expected}[f_b(E)]})}{\sum_{b \in D} d(b)}
\] (5.23)

\[
SI(E) = \frac{\sum_{a \in S} d(a) + \log(\frac{\text{Observed}[f_a(E)]}{\text{Expected}[f_a(E)]})}{\sum_{a \in S} d(a)} - \frac{\sum_{b \in D} d(b) + \log(\frac{\text{Observed}[f_b(E)]}{\text{Expected}[f_b(E)]})}{\sum_{b \in D} d(b)}
\] (5.24)

\[
SI(E) = \frac{\sum_{a \in S} d(a)(-\log_{10} h_a(E))}{\sum_{a \in S} d(a)} - \frac{\sum_{b \in D} d(b)(-\log_{10} h_b(E))}{\sum_{b \in D} d(b)}
\] (5.25)

\[
SI(E) = \frac{\sum_{a \in S} d(a) + (-\log_{10} h_a(E))}{\sum_{a \in S} d(a)} - \frac{\sum_{b \in D} d(b) + (-\log_{10} h_b(E))}{\sum_{b \in D} d(b)}
\] (5.26)

Here \(d(a)\) is still the cell diversity score associated with stemness module \(a\), while \(d(b)\) is the cell diversity score associated with differentiation module \(b\). \(f_a(E)\) is the fraction of upregulated genes in module \(a\) in experiment \(E\).

The log-ratio scores observed in Equations 5.23 and 5.24 are estimated based on the observed fraction of upregulated genes in the module and the expected fraction as calculated from the hypergeometric distribution. Finally, \(h_a(E)\) corresponds to the p-value of the gene overlap between the module \(a\) and the set of upregulated genes in the new experiment gene signature \(E\) as estimated using the hypergeometric distribution.
I tested these four different scores using three different feature types: homolog modules-only, functional modules-only and a combined feature set for a total of twelve different scoring evaluations. I measured the accuracy of each stemness index scoring method using a precision-recall curve based on a sweep of all possible cutoffs \([-10...10]\)] (Figure 5.14).

The results indicated that functional families perform globally worse than the homolog-based and combined-based feature sets regardless of the choice of scoring. The combined feature set compared better to the homolog-only based features, even though on a score-by-score basis homolog-based features performed best (Figure 5.14).

I also observed a marked improvement in the accuracy of the log-ratio method over the p-value method. The method used for combining module score elements – multiplication versus sum – was not as crucial, though the multiplication method often showed higher accuracy. For the log-ratio activation measure method that used combined features, the accuracy of the “mult” method was noticeably higher than the “sum” method (Figure 5.14).

All homolog-based feature scores showed a significant improvement over the initial weighted binary or fractional scoring methods. This marked improvement in scores was most likely due to the incorporation of information about the expected activation of individual stemness modules, as opposed to the raw numbers of upregulated genes used as a measure of gene set (module) activation in the initial more naive approaches. Additionally, it is possible that the differentiation modules provided additional independent information from the stemness modules, which improved the predictive accuracy.
Based on these results and observations, I selected the multiplicative-based log-ratio method with a homolog-based stemness and differentiation module feature set as the final method to use for measuring stemness index scores. All further analyses use Equation 5.27.

\[
SI(E) = \frac{\sum_{a \in S} d(a) \log(\frac{\text{Observed}[f_a(E)]}{\text{Expected}[f_a(E)]})}{\sum_{a \in S} d(a)} - \frac{\sum_{b \in D} d(b) \log(\frac{\text{Observed}[f_b(E)]}{\text{Expected}[f_b(E)]})}{\sum_{b \in D} d(b)}
\]  

(5.27)

To assess if the stemness and differentiation homolog modules were truly predictive, I compared the performance of the homolog feature set with the average performance of 100 randomly selected homolog feature sets of equal size using the newly defined stemness index score. The stemness and differentiation module set of features performed significantly better than random features \(t=4.3182\) \(p\)-value = 1.763e-05; paired Student t-test), as shown in Figure 5.15.

Finally, using the final stemness index scoring method I compared the pooled stemness index scores of all stem cell and differentiated cell experiments in the cross-validation sets from the mouse compendium. The stem cell experiments in the compendium had significantly higher stemness indices than the differentiated cell experiments \(t=9.8366;\) \(p\)-value < 2.2e-16; Welch two-sample t-test; Figure 5.16). These results indicated the stemness index score method could successfully distinguish between stem cell and differentiated cell experiments, which opens up some exciting application possibilities, such as the evaluation of the self-renewal potential of any cell based on its gene expression signature. Some of the possible applications are discussed further in
Figure 5.14: Precision-recall comparison of twelve stemness index scores, based on homolog-only (red), functional-only (blue) and combined (green) features. X-axis measures the recall associated with each method, while the y-axis measure the precision of each method. The most accurate method should be approximately in the top right-hand-side corner. The comparison between the twelve stemness index scoring measures suggests that the multiplicative-based log-ratio method, based on a homolog-based feature set (red dashed line) has the highest accuracy.
Figure 5.15: Precision-recall comparison of the real stemness and differentiation features to 100 randomly selected feature sets. The red line indicates the performance of the real feature set of stemness and differentiation homolog modules, while the black dashed line shows the average performance of 100 random homolog feature sets of the same size as the original feature set. The real stemness and differentiation features perform significantly better than the average random feature sets.
Chapter 8.

5.10 Summary

This chapter gave a detailed description of the SMA method designed to identify and measure stemness mechanisms in different stem cells types. Section 5.1 provided a brief overview of the method. The microarray expression data used in these studies was summarized in Section 5.2. Section 5.3 described the generation of functional and evolutionarily-related gene modules. Section 5.4 outlined the general form of the recurrence score, as well as the methods used to select the final parameters of the score. Section 5.5 introduced the cell-type and gene-based diversity scores used later for module type classification. Section 5.6 described the specificity score used to distinguish between modules used primarily by stem cells and modules used by differentiated cells as well. Section 5.7 used these three scoring measures to classify modules into different pattern types. Section 5.8 briefly discussed the methodology for the inference of stemness “off” modules, along with the stemness “on” modules. The last section, Section 5.9 outlined various scoring measures that assess how stem cell-like a gene signature is.
Figure 5.16: A box-plot comparison of the stemness indices of all stem cell and differentiated cell experiments in the mouse stem cell compendium, as defined by the cross-validation setup. Stem cell signatures (right side) in the compendium show significantly higher stemness indices than differentiated cell signatures (left side) based on cross-validation results. X-axis shows the two input types, stem cell and differentiated cell signatures, while the y-axis shows the stemness index (SI) score. The thick bands in the middle of each box represent the median values, while the lower and upper ends of each box represent the 25th and 75th percentiles.
Chapter 6

Stemness mechanisms in mouse stem cells

The goal of this chapter is to address the first central question of this dissertation: do functional redundancy and tissue-specific expression mask the common stem cell mechanisms? The chapter presents and summarizes the results of the application of the SMA method to a large mouse stem cell compendium of gene expression data, collected as part of the study. Section 6.1 describes the selection of recurrently upregulated modules from the mouse data set compendium, while Section 6.2 presents the selection of cell diverse modules. Section 6.3 summarizes the classification of all significantly recurrent mouse modules into the different pattern types introduced in the previous chapter. Section 6.4 discusses some of the most interesting stemness modules identified by the stemness meta-analysis method. Section 6.5 presents a comparison of the results of the SMA method to its most similar other study in the literature.
6.1 Identification of recurrent modules from mouse dataset compendium

I tested a collection of diverse modules of functionally- and evolutionarily-related genes. All homolog and functional modules were defined and processed as described in Section 5.3. After neighbor expansion, I identified 9,908 homolog families, comprised of 4,653 mutually-exclusive homolog groups (with two or more gene members) and 5,255 biological singletons (genes without a close homolog in the mouse genome), as well as 611 candidate functional gene modules.

I applied the recurrence, diversity and specificity scoring measures to the entire collection of mouse gene modules. For each type of score, I estimated the false-discovery rate using the permutation analysis described in Section 5.4. I found a significant shift in recurrence scores between the gene modules and the negative controls, which consisted of sets of randomly grouped genes with the same size distribution as the real gene modules (t = 7.2669, p-value = 3.498e−13; Figure 6.1).

The recurrence analysis showed that at a 5% FDR cutoff 266 homolog modules (Figure 6.2) and 94 of the functional modules (Figure 6.3) were coordinately upregulated across stem cell experiments. This is consistent with the existence of a set of homolog families with genes coordinately upregulated in stem cells.
Figure 6.1: (A.) A representative recurrence score distribution of all 1098 modules of size 3 shown in black – 1091 homolog modules and 7 functional modules – compared to randomized modules shown in yellow, based on 1000 permutations of the original data, indicates a significant shift in recurrence scores. X-axis shows the recurrent upregulation score, while the y-axis shows the number of modules in each bin. The recurrence score significance cutoff is 1.4. (B) The same recurrence score distribution is shown on a log scale. Empty bins have been assigned a floor value of 0.00001 to facilitate log scale plotting.
Figure 6.2: Selection of significant recurrently upregulated mouse homolog families. X-axis corresponds to the size-dependent recurrence score, while the y-axis show the false discovery rate on a log scale. Each color represents the FDR curve associated with a different module size. The value under each colored arrow represents the number of upregulated homolog families of that size that passed the recurrence cutoff for that size, shown in Table 5.11. FDR cutoff used to identify significantly recurrent modules was 5%.

Figure 6.3: Distribution of the modules of each input represented as significantly recurrently upregulated (yellow). Each pie chart examines only the non-singleton modules (modules with more than one gene member). Non-recurrent homolog modules are shown in blue, while non-recurrent functional modules are shown in red.
6.1.1 Recurrent module swap control

“Housekeeping” gene modules could exhibit broad upregulation across a panel of unrelated cells completely independent of stem cell mechanisms. Thus, it was possible that the recurrence analysis would identify significant modules, regardless of the input data. To assess if such housekeeping processes appreciably confounded the analysis, I performed a “swap” experiment. The purpose of this experiment was to test whether functionally similar cells, such as stem cells, coordinately upregulate genes within gene modules to a greater extent than functionally different populations, such as differentiated cells.

For most experiments included in the mouse compendium, I had two input lists of differentially expressed genes. One consisted of the genes upregulated in the stem cell and the other consisted of the genes downregulated in the stem cell (upregulated in the differentiated cell). To identify significant recurrently upregulated modules in stem cells, I usually used only the sets of upregulated genes in the stem cell.

The “swap” experiment can be described as follows: I replaced the lists used as input to the recurrence analysis, such that each list of upregulated genes in the stem cell was swapped with its corresponding list of genes downregulated in the stem cell. This swap was performed for an increasing number of experiments until all input lists had been replaced with their counterparts of downregulated genes. At each swap level, the swap was performed ten different times, such that the subset of experiments that were chosen to be swapped was randomly selected. For example, if the lists of upregulated
genes in five different stem cell populations had to be swapped, these five lists were randomly selected from the entire input set exactly ten times.

If the recurrence of modules was more tightly related to the functional similarity of the stem cells, I would expect that as I replaced the input lists of genes upregulated in the stem cells with the genes upregulated in the differentiated cells, I would identify fewer recurrently upregulated modules.

To test this expectation, I measured the FDR for functional and homolog modules separately and evaluated it as a function of the number of swaps. The cutoffs used to identify recurrent modules in each context were based on stem cell-only input (zero swaps).

Homolog modules showed a stronger swap trend than functional modules. The lowest false discovery rate of shared homolog families (FDR=6.8%±0.0006) was achieved when no swaps were introduced, indicating that stem cells share homolog families to a significantly greater extent than functionally unrelated cell populations (Figure 6.4). The functional gene families also showed their lowest false discovery rate in the datasets associated with no gene list swaps.

The FDR associated with the datasets comprised entirely of stem cells was slightly above the expected 5% (FDR=6.8% ± 0.0006) most likely due to the fact that in the swap test, false discovery rates were calculated from a smaller number of random permutations (100, instead of 1,000) than in the original analysis used to select the recurrence score cutoffs. Even though the higher number of permutations could have given the more accurate FDR level, the smaller number of permutations was used to
reduce the computational time.

6.1.2 Cultured cell bias

Because many of the populations were derived from cultured stem cells, I measured the extent to which the results might be due to the comparison of cultured cells rather than stem cells. The cultured cells included all neural and embryonic stem cell sources, as well as a couple of other stem cell types, such as spermatogonial and liver stem cells. To get a conservative upper bound on the degree of influence of highly proliferative cultured cells on the results, I excluded all data from cultured cell populations (27/49 data sets) and recalculated recurrence.

I scored the recurrence of modules using only the 22 individual primary cell populations and identified 112 modules. Similarly, using only the 27 cultured cell populations, I identified 177 modules. To estimate the extent to which the results reflect signatures from only cultured cells, I compared the modules derived from the primary, cultured, and compendium analyses (Figure 6.5). I found that most of the modules recovered in the sub-populations were also detected when the compendium was analyzed - 76 out of 112 primary modules and 139 out of 177 cultured modules.

The results show that over half of the modules (143 out of 266; 54%) identified from the compendium are due to the inclusion of primary cells (moon-shaped area in Figure 6.5). The signal in recurrent modules that did not have strong primary cell contribution was not necessarily associated with the cultured status of the cells, as the cultured cell data included all embryonic and neural stem cells, so the source of the
Figure 6.4: Swap control of mouse stem cell data. X-axis shows the number of swapped input lists of upregulated genes, where the values range from 0 swapped experiments (stem-cell-only input) to 40 (differentiated-cell-only input). Y-axis shows the false discovery rate (FDR) as a function of the number of swapped experiments. For each bar, the average FDR across ten experiments is plotted, while the error bars represent the standard error from the ten FDR summary measurements. The swap control experiment suggested that stem cells share homolog (red) and functional (blue) modules to a significantly greater extent than functionally unrelated cells (light pink: homologs; light blue: functional modules). Homolog families showed the lowest FDR and a stronger trend than functional modules, based on the more contrasting results between stem cell and functionally unrelated cells.
signal could also be stem cell-associated.

These results show that the significant recurrent modules identified by the comprehensive meta-analysis are not likely to be heavily influenced by the over-representation of cultured stem cell types in the compendium.

### 6.2 Selection of significant of cell-diversity scores

Once recurrently upregulated modules were selected, I applied the cell type-diversity measure to all significantly recurrent mouse modules. The significance cutoff selection method was identical to the recurrence scoring significance evaluation and was
already described in Section 5.5.3. I used this methodology to select the final cell-
diversity score cutoff of 2.5 ($max = 3.6$) at a false discovery rate cutoff of 5%. At this
cutoff, 114 of the 266 significantly recurrent homolog families were identified as cell-type
diverse (Figure 6.6).
Figure 6.6: Selection of significant cell-type diverse modules in the mouse stem cell compendium. (A.) The cutoff was selected as the 5% FDR cutoff score associated with the weighted average of the FDR curves for all family sizes. Each color represents the FDR curve associated with a different module size. X-axis represents the cell diversity score, while the y-axis shows the FDR in log scale. To facilitate log plotting, a floor value of 0.0001 is selected for all entries that would be otherwise 0. (B.) At the 5% FDR cutoff, 114 recurrent homolog families (red) passed the criteria and were labeled as cell-type diverse modules.
6.3 Classification of modules using diversity and specificity scoring

Using the diversity and specificity scoring schemes, I classified the significant recurrent evolutionary and functional modules into the classes described in Section 5.7.

6.3.1 Single-gene stemness

To distinguish between the contribution of the modules and the contribution of individual genes, I first, however, used the SMA method on single-genes-only input, i.e. every gene was treated as a module. Consistent with the founder studies, few genes were upregulated in common across different stem cells. In total, 38 genes had significant recurrence, cell diversity and specificity scores and I refer to them as stemness genes. Examples of stemness genes were Mcm2, Mcm4, Pcna, Set, Cdt1 and several cyclin genes. Many of the identified stemness genes have been implicated primarily in roles associated with replication and the replication fork, though they could have other unknown roles in the cell as well. The full list of stemness genes is shown in Table 6.1.

6.3.2 Module-level stemness

At the evolutionary and functional module level I identified 124 all-for-all (AFA) stemness modules and 38 one-for-all (OFA) stemness modules (Table 6.2; Figure 6.7). Of the OFA stemness homolog modules, 4 were trivial as they were among the single stemness genes identified earlier. Homolog modules were enriched among the
### Table 6.1: A summary of the 38 stemness genes identified by the single-gene SMA method is shown along with the functional categories significantly enriched in the stemness genes. The first column gives the name of the functional category. The second column shows the names of the stemness genes that fall in that functional category. Genes that have been associated with more than one significant functional categories are italicized. Genes placed in the “Other” category are not identified with any significant functional group. The third column shows the p-value associated with the significance of the overlap, as measured by the hypergeometric distribution, between the set of all 38 stemness genes and the functional category represented in that row. If the p-value > 0.05, no value is shown.

<table>
<thead>
<tr>
<th>Functional association</th>
<th>Stemness genes</th>
<th>Enrichment p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA replication</td>
<td>Pcna, Cdt1, Orc11, Mcm2, Mcm4, Mcm5, Cdc6, Rrm2</td>
<td>1.243713e − 05</td>
</tr>
<tr>
<td>Cell division</td>
<td>Top2a, Ccnd2, Mcm5, Cks2, Bub1, Cdc6, Hells, Ruvbl1</td>
<td>0.0008</td>
</tr>
<tr>
<td>Chaperonin</td>
<td>Cct3, Cct5, Cct8</td>
<td>—</td>
</tr>
<tr>
<td>Other</td>
<td>Kpna2, Ncl, Nap111, Dph5, Ttk, Coll1a1, Impdh2, Ipo5, Shmt1, Depdc6, Set, Fignl1, Dnahc11, Shroom3, Prps1, Hnrupa2b1, Sfrs3, Dtymk, Csp2, Eya2, Fbl</td>
<td>—</td>
</tr>
</tbody>
</table>
identified AFA and OFA patterns.

Interestingly, approximately 70% (80/110) of the stemness AFA and OFA homolog modules did not contain a single recurrent stemness gene. Therefore, the diversity of these homolog modules derives mainly from the complementary pattern of upregulation across different member genes and could not have been identified using single gene-level analysis.

In summary, 78 stemness homolog AFAs (HAFA; Table B.1) and 46 functional stemness AFAs (FAFA; Table B.2) were identified, as well as 25 stemness homolog OFA modules (HOFA; Table B.3) and 13 stemness functional OFA modules (FOFA; Table B.4). These results indicate an approximately 3-fold enrichment of modules (both homolog and functional) that upregulate a high number of their member genes (AFA modules) over modules with a lower gene usage (OFA modules). For comparative purposes, if we select for non-stemness (tissue-specific) patterns, there is only 1.2-fold enrichment of modules that use a high number of their member genes over modules that use a low number of their member genes.
<table>
<thead>
<tr>
<th>Class</th>
<th>Cell diversity</th>
<th>Gene diversity</th>
<th>Specificity</th>
<th>Homolog modules</th>
<th>Functional modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFA</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>78</td>
<td>46</td>
</tr>
<tr>
<td>OFA</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>25(21)</td>
<td>13</td>
</tr>
<tr>
<td>AFO</td>
<td>-</td>
<td>+</td>
<td>+/-</td>
<td>84</td>
<td>17</td>
</tr>
<tr>
<td>OFO</td>
<td>-</td>
<td>-</td>
<td>+/-</td>
<td>68(42)</td>
<td>15</td>
</tr>
<tr>
<td>CM</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>CG</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>266(236)</td>
<td>94</td>
</tr>
</tbody>
</table>

Table 6.2: Summary of the classification and distribution of all recurrently upregulated gene modules in the mouse stem cell compendium. The stemness modules (AFA and OFA pattern classes) identified by the SMA method are shown in bold.
Figure 6.7: Global overview of classes of functional and homolog recurrent modules. Red and green demarcate all stemness homolog and functional modules. All recurrent modules that did not pass the criteria for stemness are shown in grey.
I observed many families with a high level of interchangeability between individual gene members, which suggests the dispensability of individual genes. One clear example of tissue specificity among the stemness AFA homolog families came from the p53 family of genes, specifically p53 (Trp53 in Figure 6.8), p63 (Trp63) and p73 (Trp73). p63 is necessary primarily for the maintenance of epithelial stem cells and the stemness meta-analysis showed a clear separation of the expression of this gene in breast, gastric and intestinal stem cells (Figure 6.8), while p53 – its well-known paralog – was upregulated in most other stem cell types, including hematopoietic, embryonic and neural stem cells.

The stemness analysis results indicate that some homolog families – Myb (Figure 6.9) and Rbp (Figure 6.10) families – show a clear alternation in gene usage between different stem cell types, while other groups can use a variable set of genes within the different types of stem cells.

The mechanisms of specialization to these different stem cells are particularly interesting. Specifically, some families may have developed a mostly exclusive cell type specificity of individual genes, which we can trace through the evolution of the homolog family. Other groups may, however, allow for a more stochastic nature of homolog partner gene use in the cell, which may be advantageous for the maintenance of the stem cell state. Understanding the differences between these module types may also allow us to better predict the contribution of individual genes to self-renewal.
Figure 6.8: Stem-cell-only expression pattern of the p53 tumor suppressor gene family – p53 (Trp53), p63 (Trp63), and p73 (Trp73). The pattern shows the specialization of p53 and p63 to different cell types. Each column represents one of the twelve stem cell types. Each row represents a different module member gene. The value for each gene in a given cell type is calculated as the average number of experiments of that stem cell type that measured the gene as upregulated, ranging from 0 (the gene was not upregulated in a single experiment) to 1 (the gene was upregulated in all experiments of that cell type). Genes that have not been tested in a given cell type are shown in grey.
Figure 6.9: Stem-cell-only expression pattern of Myb gene family – a-myb (Mybl1), b-myb (Mybl2), and c-myb (Myb). Each column represents one of the twelve stem cell types. Each row represents a different module member gene. The value for each gene in a given cell type is calculated as the average number of experiments of that stem cell type that measured the gene as upregulated, ranging from 0 (the gene was not upregulated in a single experiment) to 1 (the gene was upregulated in all experiments of that cell type). Genes that have not been tested in a given cell type are shown in grey. The family shows a very diverse pattern of upregulation across different family members.
Figure 6.10: Stem-cell-only expression pattern of Rbp gene family – Rbp1, Rbp2, and Rbp7. Each column represents one of the twelve stem cell types. Each row represents a different module member gene. The value for each gene in a given cell type is calculated as the average number of experiments of that stem cell type that measured the gene as upregulated, ranging from 0 (the gene was not upregulated in a single experiment) to 1 (the gene was upregulated in all experiments of that cell type).
6.4 Stemness modules

Families whose member genes are involved in the maintenance of the stem cell state across a large number of stem cell types are particularly biologically interesting, because they can be potentially used as markers for stem cell types that are not well understood. Alternatively, these genes could also be candidates for pluripotency or multipotency induction in mature cells of various cell types.

I primarily focused on the 103 significant stemness (AFA and OFA; Tables B.1 and B.3) homolog families that showed the highest level of stem cell type diversity. Functional enrichment analysis of the stemness homolog modules showed a wide variety of functional categories varying from Wnt pathway signaling to chromatin assembly, phosphatase activity, ligase and ATPase activities and DNA repair (Figure 6.11; upper panel). I also identified stemness functional modules associated with imprinting, chromatin-dependent silencing, heterochromatin and the nuclear lamina, consistent with a wide-spread suppression of many lineage-associated genes before differentiation (Figure 6.11; lower panel).

A few themes emerged pertaining to the stemness families. These themes provide an understanding of how the cell may achieve the balance between quiescence, proliferation, apoptosis, and differentiation. I observed several proto-oncogene families, balanced by the expression of more than a few tumor suppressor factor families; signaling pathways known to increase self-renewal along with proteins known to downregulate signaling and bring back the cell to quiescence. Other regulatory molecules included
Figure 6.11: Functional categories represented in the stemness homolog (top panel) and functional (bottom panel) gene modules. Each row represents a different stemness module. The value for each module in a given cell type is calculated as the fraction of upregulated genes in the module in the stem cell type. The value could range from 0 (no genes upregulated in the stem cell type) to 1 (all genes upregulated). The categories shown to the right of the upper panel heatmap consist of functional categories significantly enriched in the stemness homolog modules, as measured by the hypergeometric distribution. The categories shown to the right of the lower panel heatmap consist of representative names of functional modules.
chromatin remodeler families along with lineage-specific inhibitors. These molecules were supported by chaperone proteins and signal transduction proteins. As the niche input is indispensable to a stem cell, I also found common adhesion molecules.

In the next few paragraphs, I briefly discuss some of the most interesting examples of stemness families and summarize the current knowledge on their role in stem cell biology. The reader should note that several families (Itga, Frizzled, TCF/LEF, and Chd/Smarc) are left out of the discussion in this chapter, as they also make an appearance in the human stemness analysis, so they are discussed in detail in the next chapter.

### 6.4.1 Oncogenes: Myb family

One of the highest scoring stemness families was the Myb family of oncogenes: a-myb (Myb1 in Figure 6.9), b-myb (Myb2) and c-myb (Myb; Figure 6.9). I found a very diverse gene expression pattern in this family among the different stem cell types. Previously, c-myb has been suggested to control hematopoietic proliferation and differentiation [149]. Recently, it was also implicated as a potential master regulator of differentiation, as RNAi-induced silencing of this gene in a human leukemic cell line mimicked very closely the effects of the application of a differentiation inducing drug [169]. In our murine stemness analysis, c-myb was activated not only in the hematopoietic system, but also in neural, embryonic, intestinal and retinal stem cells. A-myb complemented the expression pattern of its partner genes by showing significant upregulation in gastric stem cells, while b-myb was upregulated in trophoblast stem
cells. The neural and embryonic stem cell types used all members of this proto-oncogene family.

6.4.2 Tumor suppressor factors: Sfrp family

Aside from the p53 family, other putative tumor suppressor families could also play important roles in stem cell regulation. Specifically, while the self-renewal-associated Wnt pathway had a significant presence in the stemness family set with two of its most important components – the Frizzled family of receptors and the TCF/LEF family of transcription factors (effectors of β-catenin accumulation), one of the largest families of Wnt inhibitors, the secreted Frizzled-related protein (Sfrp) family, was also classified as a stemness family.

In human, Sfrp has five protein members, but three of them – Sfrp1, Sfrp2, and Sfrp5 – represent an independent subfamily from Sfrp3 and Sfrp4 with different ligands [153]. Interestingly, the stemness analysis recognized Sfrp1, Sfrp2 and Sfrp5 as a separate, putative stemness family. The stemness upregulation pattern (Figure 6.12) showed a wide use of all family members in different stem cell types.

Sfrp genes could function in several different ways – they could either interact with the Wnt proteins directly, or they could modulate each other or various other receptors and inhibit BMP signaling [22]. Sfrp2 has also been shown to interact with the fibronectin integrin α5β2 receptor complex and regulate cell adhesion [22]. One interesting model in intestinal stem cells suggests that Sfrp genes can function as part of Hedgehog signaling, where in normal intestinal differentiated cells, Hedgehog is active
Figure 6.12: Secreted Frizzled-related protein (Sfrp) family expression in stem cells. The Sfrp family is one of the major inhibitors of the spread of Wnt signaling and shows a very diverse pattern of upregulation in different stem cells. Each column represents one of the twelve stem cell types. Each row represents a different module member gene. The value for each gene in a given cell type is calculated as the average number of experiments of that stem cell type that measured the gene as upregulated, ranging from 0 (the gene was not upregulated in a single experiment) to 1 (the gene was upregulated in all experiments of that cell type).
Figure 6.13: Sfrp1 model of regulation in intestinal stem cells. The model is adapted from a paper by Katoh and Katoh [89]. Dark blue indicates the stem cells located near the stem cell niche, the lighter blue shows intestinal progenitor cells, while the lightest blue highlights the differentiated cells. Sfrp1 can inhibit the spread of Wnt signaling from the stem cells to the differentiated cells.

and acts through Sfrp1, which regulates Wnt signaling to ensure that Wnt will not spread from the intestinal stem cells at the base of the crypt niche to the differentiated cells further up [89] (Figure 6.13).

6.4.3 NM23 family

Another highly scoring stemness family of proteins consisted of the NM23 group of homologs, which play functional roles in differentiation and tumorigenesis [125, 138]. This homolog family was represented by three genes, Nme1, Nme2, and Nme4, and has been previously shown to have a tissue-specific and differentiation-specific manner of expression. In particular, the human NM23 family has been implicated in negative
regulation of stem cell differentiation – NME1 negatively regulates growth factors and NME2 has been implicated as a direct activator of c-myc [138]. The stemness analysis results suggest that this family may be involved in the regulation of most stem cell types, where Nme2 and Nme4 are the predominantly active genes, but Nme1 and Nme3 also show upregulation in embryonic and spermatogonial stem cell types (Figure 6.14).
Figure 6.14: Non-metastatic expressed (Nme) family expression in stem cells represents another example of a tumor suppressor family with a role in stem cell fate. Each column represents one of the twelve stem cell types. Each row represents a different module member gene. The value for each gene in a given cell type is calculated as the average number of experiments of that stem cell type that measured the gene as upregulated, ranging from 0 (the gene was not upregulated in a single experiment) to 1 (the gene was upregulated in all experiments of that cell type).
6.4.4 Chaperone roles: Heat shock (Hspa) and importin families

One very interesting stemness family was the heat shock protein (Hspa/Hsp70) module. Heat shock proteins belong to a set of chaperones and co-chaperones whose cellular role is to aid the folding process of newly formed proteins through conformational and other changes. Because of this essential role, Hsp proteins are expressed generally in all cells, however they are highly induced in cells undergoing a stress response. The role of these proteins has not been heavily studied in stem cells, but several studies have suggested that these chaperone proteins may be involved in stem cell self-renewal, as their expression significantly reduces with differentiation [139].

Many different classes of heat shock proteins exist, such as Hsp90, Hsp70, Hsp60 and others, but the stemness meta-analysis identified only the Hsp70/Hspa module of proteins – Hspa1a, Hspa1b, Hspa1l, Hspa2, Hspa4, Hspa4l, Hspa5, Hspa8, Hspa9, Stch (Hspa13), Hspa14, Hsph1, and Hyou1(Hsph4) – as commonly upregulated across different stem cell types (Figure 6.15). This result is consistent with a study that shows the active upregulation of many Hsp70 member genes in embryonic, neural and mesenchymal cells [12].

One model of regulation associated with Hsp genes (though not directly Hsp70-associated) starts with the arrival of a signaling molecule in the cell. After its activation, the signaling molecule binds the chaperone and co-chaperone complex. Subsequently the entire bound complex associates with an importin molecule, which facilitates the transition across the nuclear pore [139]. Importin genes are members of another family
The wide variety of genes in this family used by stem cells points to a potential role of these genes in the control of cell fate through stress response. Each column represents one of the twelve stem cell types. Each row represents a different member gene of the Hspa/Hsp70 family of proteins. The value for each gene in a given cell type is calculated as the average number of experiments of that stem cell type that measured the gene as upregulated, ranging from 0 (the gene was not upregulated in a single experiment) to 1 (the gene was upregulated in all experiments of that cell type).
designated as a stemness module in the SMA analysis. After entry into the nucleus the complexes dissociate and the signaling molecule can proceed with its function. A direct application of this model is the activation of STAT3, which bound by Hsp90 could be transported to the nucleus, where it can upregulate Nanog, a gene essential for ESC self-renewal [139].

But why would stem cells require higher activity level of chaperone proteins? This question still remains unanswered. I speculate this phenomenon may be related to the importance and rarity of stem cells. It is possible that in rare cells, whose proper function is essential for the organism, proper folding and location of new proteins is crucial. This may not be as much the case for the more dispensable differentiated cells, which could explain the downregulation of these genes with differentiation. Alternatively, stem cells may be exposed to hypoxic conditions near the niche, which may induce a stress response. Finally, it is possible that the heat shock proteins of this family are induced only as a side effect of the general stress response, or else they may have another functional role, unrelated to their chaperone abilities.

6.4.5 Lineage-specific gene inhibition: Inhibitor of differentiation/DNA binding (Id) family

The inhibitor of differentiation (Id) set of proteins represents one of the seven sub-families of the very large and diverse helix-loop-helix (HLH) protein family [44]. The HLH family has more than 200 members genes and is involved in the regulation of development and cell fate decisions in a variety of organisms. Proteins in this large
family are characterized by their largely conserved HLH domain pattern and are known to assemble into either homo- or heterodimers [44].

Interestingly, the Id sub-family does not have the DNA binding capacity (Figure 6.16A) that characterizes many other HLH family members. Unlike most other HLH family members, which specifically bind to the E-box motif (CANNTG) on the DNA, Id family members interact directly with proteins from other sub-families – such as some regulators of differentiation like MyoD, NeuroD, and E2-2 – to prevent their access to the DNA. This binding event effectively blocks differentiation (Figure 6.16B), which also explains the origins of the Id family name.

In mammalian species, the Id family consists of four genes – Id1, Id2, Id3, Id4. Individual gene members have been shown to play significant roles in maintain-
ing long-term hematopoietic stem cell renewal, even though they are not an essential component of short-term engraftment [135]. Most recently, members of this family have been implicated in neural stem cell self-renewal through their inhibition of the NeuroD transcription factor [84].

6.5 Comparison with other global stemness methods

All observations discussed up to this point render the SMA approach particularly relevant to the study of stemness. I have identified highly pertinent putative stemness modules that can be used to form hypotheses about the roles of individual genes in specific stem cell types and guide their further functional study. The method does not require the use of many homologous genes in stem cells; single genes are also detected. But where does this approach stand with respect to other stemness-related methods?

The SMA method is complementary to other large-scale efforts to catalog different stem cells and understand the mechanisms of pluri- and multi-potency maintenance, such as the PluriNet [117] by Muller et al. and the stem cell module map [194] by Wong et al. The SMA method may also be more encompassing than both of these methods, since homology provides orthogonal information to KEGG pathways and GO functional gene sets (Figure 6.17) used as the basis for the Wong et al. module definition in the stem cell module map [194].

I directly tested the gene modules in the module map using the stemness meta-
A. Highly cell type diverse recurrent families

B. Recurrent families

C.

<table>
<thead>
<tr>
<th>Gene set size</th>
<th>Number of cell type diverse recurrent homolog families</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most significant overlapping gene set &lt; 100 genes</td>
<td>68</td>
</tr>
<tr>
<td>Most significant overlapping gene set &gt; 99 genes</td>
<td>46</td>
</tr>
</tbody>
</table>

Figure 6.17: Global similarity comparison between homolog modules and GO gene sets/KEGG pathways. (A.) The plot examines the size of the most similar pathway/gene set to each one of the highly cell-diverse recurrent homolog modules. X-axis shows the fraction of gene overlap between the homolog module and its most similar pathway. Y-axis shows the size of that pathway. The density of plotted points at a given set of coordinates can be inferred through the color and size of the point (single points: small and black; high number of points: large red sunflower-like point). (B.) Same as (A), but for all recurrent homolog modules. (C.) The table examines the number of cell-diverse recurrent homolog modules in each of two categories. The first category represents the set of modules, for which the most significant overlapping pathway/gene set has less than 100 gene members. The second category represents the set of modules, for which the most significant overlapping pathway has 100 or more gene members. Homolog modules in the second category contribute more specific information than their large overlapping pathways/gene sets.
analysis approach. Although the module map modules exhibited comparable recurrence scores, the highest scoring homolog families scored higher than any Wong et al. module (Figure 6.18), measured as a function of the deviation from the recurrence significance cutoffs. The role of c-myc was central to the findings of the module map study, but my stemness analysis also identified the myc-family – c-myc, N-myc, v-myc – as a highly relevant putative stemness “master regulator” module.

6.6 Differentiation modules

In addition to the stemness (stemness “on”) families, I also defined differentiation (stemness “off”) modules from the entire original stem cell compendium. Because the stemness index score only used homolog features, I defined only differentiation homolog families. It should be noted that the differentiation modules were derived from functionally more heterogeneous experimental data, so I expected a smaller number of differentiation-specific homolog families. The SMA method identified a total of 39 differentiation modules, summarized in Table 6.3.

The presence of a few specific homolog families warrants special mention. Both the integrin β (Itgb; 10th row in Table 6.3) and ATP-binding cassette, subtype B (Abcb; 28th row in Table 6.3) families of proteins have important roles in stem cells. Integrin β protein family members function as part of an integrin heterodimer receptor, which facilitates the communication of stem cells with their respective niche. ABC transporter proteins are actively used by stem cells to remove various toxins and drugs from the
Figure 6.18: A direct comparison of the Wong et al. “stem cell module map” modules (green) and the SMA homolog families (black). X-axis shows the deviation of the recurrence score for each module from the significance recurrence cutoff for that module size. Y-axis shows the module frequency in each bin in log scale. All empty bins have been assigned a floor value of 0.00001 to facilitate log scale plotting.
cell. Thus, it is surprising to see these two families among the differentiation modules.

While the upregulation patterns of these two families were sufficiently dominated by genes used in differentiated cells and even though these two modules did not pass the recurrence and diversity cutoffs to be defined as stemness families, they still showed upregulation of individual genes in stem cells. For example, in the ATP-binding cassette, subtype B (or Abcb) module, Abcb1b, Abcb10, and Abcb8 were almost exclusively upregulated by stem cells (Figure 6.19).

Therefore, the differentiation module results are not in conflict with the current literature. It is a bit unusual to observe members of the same heterodimer complex regulated differently – integrin α is significantly upregulated in stem cells, while integrin β is significantly downregulated, but it should be noted that the most common integrin complexes used in stem cells make use of only a minority of the members of the integrin β family, such as integrin β1.
<table>
<thead>
<tr>
<th>Differentiation module name</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gna</td>
<td>3.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Ankrd</td>
<td>2.84</td>
<td>0</td>
</tr>
<tr>
<td>Cebp</td>
<td>2.83</td>
<td>0</td>
</tr>
<tr>
<td>S100a</td>
<td>2.78</td>
<td>0</td>
</tr>
<tr>
<td>Pdlim</td>
<td>2.78</td>
<td>0.008</td>
</tr>
<tr>
<td>H2</td>
<td>2.74</td>
<td>0</td>
</tr>
<tr>
<td>Flt</td>
<td>2.74</td>
<td>0</td>
</tr>
<tr>
<td>Klf</td>
<td>2.73</td>
<td>0.0018</td>
</tr>
<tr>
<td>Bhll1b</td>
<td>2.71</td>
<td>0.0147</td>
</tr>
<tr>
<td>Itgb</td>
<td>2.70</td>
<td>0.004</td>
</tr>
<tr>
<td>Lgals</td>
<td>2.70</td>
<td>0.0027</td>
</tr>
<tr>
<td>Atp1a</td>
<td>2.69</td>
<td>0</td>
</tr>
<tr>
<td>Vil</td>
<td>2.68</td>
<td>0.0196</td>
</tr>
<tr>
<td>Tspan</td>
<td>2.67</td>
<td>0</td>
</tr>
<tr>
<td>Ces</td>
<td>2.67</td>
<td>0</td>
</tr>
<tr>
<td>Anxa</td>
<td>2.66</td>
<td>0</td>
</tr>
<tr>
<td>Soat</td>
<td>2.65</td>
<td>0.0334</td>
</tr>
<tr>
<td>Lck/Src/Fyn/Abl</td>
<td>2.65</td>
<td>0</td>
</tr>
<tr>
<td>Mef2</td>
<td>2.65</td>
<td>0.004</td>
</tr>
<tr>
<td>Flvcr</td>
<td>2.64</td>
<td>0.0334</td>
</tr>
<tr>
<td>Degs</td>
<td>2.63</td>
<td>0.0372</td>
</tr>
<tr>
<td>Cpxm</td>
<td>2.62</td>
<td>0.008</td>
</tr>
<tr>
<td>Vamp</td>
<td>2.62</td>
<td>0.0372</td>
</tr>
<tr>
<td>Ero1</td>
<td>2.62</td>
<td>0.048</td>
</tr>
<tr>
<td>Lrrc4</td>
<td>2.62</td>
<td>0.0335</td>
</tr>
<tr>
<td>Masp/C1r</td>
<td>2.59</td>
<td>0.0025</td>
</tr>
<tr>
<td>Sema6</td>
<td>2.59</td>
<td>0</td>
</tr>
<tr>
<td>Abcb</td>
<td>2.59</td>
<td>0</td>
</tr>
<tr>
<td>Itm2</td>
<td>2.57</td>
<td>0</td>
</tr>
<tr>
<td>Stgal</td>
<td>2.56</td>
<td>0</td>
</tr>
<tr>
<td>Mpp</td>
<td>2.55</td>
<td>0</td>
</tr>
<tr>
<td>Cts</td>
<td>2.55</td>
<td>0</td>
</tr>
<tr>
<td>Gabarap</td>
<td>2.55</td>
<td>0.0238</td>
</tr>
<tr>
<td>Ifi27</td>
<td>2.55</td>
<td>0.0081</td>
</tr>
<tr>
<td>Acs1</td>
<td>2.54</td>
<td>0.00355</td>
</tr>
<tr>
<td>Ppap2</td>
<td>2.54</td>
<td>0.0268</td>
</tr>
<tr>
<td>Jak</td>
<td>2.52</td>
<td>0.0456</td>
</tr>
<tr>
<td>H2-Ea/Oa/Aa</td>
<td>2.51</td>
<td>0.0055</td>
</tr>
<tr>
<td>Cpeb</td>
<td>2.50</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Table 6.3: List of all mouse differentiation homolog modules.
Figure 6.19: Upregulation pattern of the ATP-binding cassette, subtype B (Abcb) family of transporter proteins in differentiated (green) and stem (red) cells. Each column represents one of the twelve stem cell types. Each row represents a different module member gene. The value for each gene in a given cell type is calculated as the difference between the average number of experiments of that stem cell type that measured the gene as upregulated and the average number of differentiated cell experiments in that system type. The values can range from -1 (the gene was upregulated in all differentiated cell experiments of that type and no stem cell experiments of that cell type) to 1 (the gene was upregulated in all stem cell experiments of that cell type and no differentiated cell experiments of that type). Some genes, such as Abcb1b, are used primarily by stem cells, though the method identifies the family as a differentiation family.
6.7 Summary

This chapter addressed the first central question of this dissertation: do functional redundancy and tissue-specific expression mask the common stem cell mechanisms? It presented the results of the application of the SMA method to a large murine stem cell expression data compendium and showed that if we account for functional redundancy, common stemness mechanisms do emerge. Section 6.1 identified significantly recurrent reproducibly upregulated mouse homolog and functional modules. Section 6.2 described the selection of cell type diverse modules. The next section, Section 6.3, presented the classification of recurrent modules into several different pattern types. Section 6.4 summarized the biological knowledge from the literature associated with some of the most interesting stemness modules. Section 6.5 compared the stemness meta-analysis method to its most similar stemness study, while Section 6.6 presented an overview of the differentiation (stemness “off”) modules identified in the mouse stemness analysis.
Chapter 7

Stemness mechanisms in human stem cells

The goal of this chapter is to address the second central question of this dissertation: if common stem cell mechanisms exist, are they conserved between mouse and human stem cells? The chapter summarizes the results of the application of the SMA method to a human stem cell compendium and examines the conservation of stemness patterns between mouse and human. As the stemness meta-analysis setup has already been described extensively in the previous chapters, I try to avoid repetition and present only the most important summary statistics and comparisons. Section 7.1 describes the identification of recurrently upregulated human modules, compares the recurrent modules between mouse and human and explains the differences observed in the human data. Section 7.2 summarizes the selection of cell diverse modules and stemness families. Section 7.3 presents an overview of several interesting human stemness families,
while the final section – Section 7.4 – summarizes our current biological knowledge on the conserved mammalian stemness modules.

7.1 Recurrent modules in a human stem cell compendium

To identify human stemness modules, I performed the exact same steps as already described for the mouse stemness meta-analysis. I defined human homolog families and after neighbor expansion, I identified 9,081 homolog families, comprised of 4,350 mutually-exclusive homolog groups (with two or more gene members) and 4,731 biological singletons, as well as 889 non-redundant functional gene modules. The breakdown of modules was shown earlier in Tables 5.6 and 5.8.

To test the existence of stemness families in human and evaluate the conservation of stemness mechanisms between the two mammalian species, I first directly compared the two sets of predicted homolog modules from each organism. For each of the 9,081 homolog families in the human network, I mapped every gene member to its best reciprocal BLAST hit in mouse. If no such gene existed, the human gene was removed from further consideration. In the entire human homolog module network 13,711 genes had a best reciprocal hit in mouse and after the elimination of genes that could not be mapped, the orthologously defined (human-to-mouse) homolog modules were reduced to 6,755 gene families.

I evaluated the overlap between all mouse homolog modules and their corresponding best matching human modules. Homolog modules most often completely over-
lapped (Figure 7.1). A large number of the mouse homolog modules (approx. 3,000) showed no overlap with any human homolog module. These represented the mouse modules of genes that had no best reciprocal human mapping. The non-negligible peak at 0.5 consisted of families that share half of their genes between organisms. This observation is not surprising, as the mean number of genes in the mouse homolog gene set was small – close to four genes.

These results indicated that regardless of the comparison base, the fraction of gene overlap between each pair of modules was consistently high (Figure 7.1), so it is reasonable to treat any common stemness gene families between the two organisms as mammalian stemness modules.

The recurrence analysis showed that at a 5% FDR cutoff 85 homolog modules (Figure 7.2) and 34 of the functional modules were coordinately upregulated across stem cell experiments.

An initial striking difference between the recurrence results in the mouse and human stemness meta-analyses was the approximately 3-fold reduction in significant recurrently upregulated homolog modules in the human stem cell compendium as compared to mouse. To assess quantitatively this observation, I directly compared the recurrence score distributions of all human and mouse homolog families and found a significant global decrease in recurrence scores \( t = 21.579; p\text{-value} < 2.2e-16; \) one-tailed Student t-test; Figure 7.3; panel B), which could not be accounted for by a significant difference in homolog module sizes (Figure 7.3; panel A). The recurrence cutoffs used in mouse and human were also very similar (Figure 7.4), so the global difference in scores
Figure 7.1: Fraction of overlap distribution between human and mouse networks. X-axis shows the fraction of gene overlap between each mouse homolog module and its most significantly overlapping human homolog module, as measured by the hypergeometric distribution. The overlap can be shown as a function of the size of the mouse homolog module (left), or as a function of the size of the human module (right). The y-axis shows the number of homolog modules in each bin on a log scale. A large number of mouse homolog modules (approx. 3,000) show no overlap with any human homolog module. These represent the mouse modules of genes that had no best BLAST reciprocal human mapping.
Figure 7.2: Selection of significant recurrently upregulated homolog families. X-axis corresponds to the size-dependent recurrence score, while the y-axis show the false discovery rate on a log scale. Each color represents the FDR curve associated with a different module size. The value under each colored arrow represents the number of upregulated homolog families of that size that passed the recurrence cutoff for that size. FDR cutoff used to identify significantly recurrent modules was 5%.
Figure 7.3: Comparison of the global distributions of (A) homolog family sizes and (B) recurrence scores in human and mouse. (A.) X-axis shows the sizes of mouse (green) and human (red) homolog modules on a log scale. Y-axis shows the frequency of each size module on a log scale. (B.) X-axis shows the recurrence score for each mouse (green) and human (red) homolog module. Y-axis shows the frequency of modules with each recurrence score. The black arrow points to the significant (p-value shown in the figure) global decrease in the recurrence scores of the human homolog modules, as compared to the recurrence scores of the mouse homolog module.

between the two organisms was unlikely to be a function of a different scale.

One possible explanation for the reduction in reproducibility and recurrence is the higher heterogeneity of the human stem cell data. While the study of human embryonic stem cells has faced debate, much work has been done on understanding pluripotency mechanisms using both embryonic stem cells and, more recently, induced pluripotent cells in human. Experiments on human non-cultured adult stem cells, however, especially ones associated with intestinal, gastric, spermatogonial, liver and other
Figure 7.4: Correlation between the recurrence cutoffs used for every module size in common between mouse and human. X-axis shows the recurrence cutoff scale for mouse modules, while the y-axis shows the recurrence cutoff scale for human modules. Each point represents a different module size represented in both the human and the mouse input data. The \( y=x \) line is shown as a black solid line.
stem cells is much harder to perform than in mouse, so little data is available for many of these stem cell types. This experimental bias is obvious in the human stem cell compendium used in the human stemness meta-analysis. Because of these difficulties, the isolation of pure stem cell populations has been hindered and many of the input profiling data could represent more functionally diverse, but stem cell-enriched populations. The heterogeneity in the purity of the populations may explain the difference in module recurrence.

This conclusion is not at odds with the perfect clique formation of human stem cell experiments from the same stem cell type in the discovery of “replicate” sets, shown earlier in Figure 5.6. The high similarity between experiments within the few large “replicate” sets could be explained by technical issues. Experiments from the same study within the same cell type, but with different differentiation fates clustered the closest together. Other experiments that clustered very closely in a “replicate” set represented populations derived from labs with similar protocols. While these two examples show a high within-protocol reproducibility, they do not guarantee similarities with other stem cell experiments derived using different protocols.

Other explanations for the higher heterogeneity in the human data include the more diverse genetic background of humans as compared to mice, as well as transcriptional differences in the stem cell populations associated with the age and health of the humans from which the populations were derived.

Finally, yet another explanation is that there is a true difference between the mechanisms that guide mouse and human stem cell pluri- and multipotency. However,
some evidence already suggests that at least some mechanisms should be conserved. For example, the initial four transcription factors successfully used to induce the transformation of fibroblasts into induced pluripotent stem (iPS) cells—Sox2, Oct3/4, Myc and Klf4—were the same in both mammalian species [170, 171]. Even though the murine stemness meta-analysis identified these transcription factors as significantly recurrently upregulated in stem cells, the human SMA did not recognize a single one of the factors as recurrently upregulated.

To test the heterogeneity of the human stem cell data, I made use of the same “swap” analysis that I introduced earlier in the mouse stemness analysis. I swapped the lists used as input to the recurrence analysis, such that each list of differentially upregulated genes in a stem cell was replaced with its corresponding list of genes differentially downregulated in the stem cell. This swap was performed for an increasing number of experiments until all input lists had been replaced with their counterparts of downregulated genes. The only difference from the mouse test was in the number of swaps performed at each level of increase; in human, I performed only five swaps, while in mouse I used ten. Specifically, at each swap level, the swap was performed five different times, such that the subset of experiments that were chosen to be swapped was randomly selected. The baseline cutoffs used to assess the false discovery rates were used directly as derived from the stem-cell-only input data.

If the recurrence of modules was more tightly related to the functional similarity of the stem cells, I would expect that as I replaced the input lists of genes upregulated in the stem cells with the genes upregulated in the differentiated cells, I would identify
fewer recurrently upregulated modules and higher FDR levels. Also, if the human stem cell populations were more heterogeneous than their mouse counterparts, the distinction between the FDR associated with the stem cell-only data and the FDRs of the progressively swapped datasets would not be as clear as it was in mouse. Intuitively, equally functionally dissimilar cells should generate approximately equal false discovery rates. This result would suggest that the stem cell data resembles less functionally similar populations, so the heterogeneity of the input stem cell populations would be a good explanation.

As already mentioned, to reduce computational time I limited the number of swaps to five at each level of increase, still allowing the estimation of a standard error. The results (Figure 7.5) confirmed the prediction: while the false discovery rate associated with the stem-cell-only dataset was still the lowest, the false discovery rates for the more heterogeneous sets were following more closely the stem-cell-only FDR than previously observed in the mouse analysis.

7.2 Cell type diversity assessment and classification of recurrent modules

To identify significantly cell type diverse modules, I used the methods described in Chapter 5 and chose a cutoff of 2.15 at a 5% FDR (Figure 7.6).

At this cutoff, 22 evolutionary families and 10 functional modules satisfied the stringent criteria. None of these modules were also significantly recurrent in different-
Figure 7.5: Swap analysis of human recurrently upregulated homolog families. X-axis shows the number of swapped input lists of upregulated genes, where the values range from 0 swapped experiments (stem-cell-only input; marked in red) to 38 (differentiated-cell-only input). Y-axis shows the false discovery rate (FDR) as a function of the number of swapped experiments. For each bar, the average FDR across five experiments is plotted and the error bars represent the standard error from the five FDR summary measurements.
Figure 7.6: Selection of significant cell-type diverse modules in the human stem cell compendium. The cutoff (2.15) was selected as the 5% FDR cutoff score associated with the weighted average of the FDR curves for all family sizes. Each color represents the FDR curve associated with a different module size, as shown in the legend. X-axis represents the cell diversity score, while the y-axis shows the FDR in log scale. To facilitate log plotting, a floor value of 0.0001 is selected for all entries that would be otherwise 0.
tiated cells, so this set of families represented the final group of stemness modules. A summary of the distribution of modules into individual classes is shown in Table 7.1 and the list of all stemness families is shown in Table 7.2. I did not observe any significant deviations in the class membership module distribution between mouse and human. As in the mouse stemness analysis, AFA modules were at least four-fold overrepresented than their partner OFA modules in both the human homolog and functional recurrent gene modules. The human stemness analysis did not generate any constitutive module sets or gene sets, although given the small number of recurrently upregulated modules altogether, this result is not entirely surprising.

Next, I compared the conservation of the stemness homolog families between mouse and human – approximately 6,700 modules were shared between the two species – and found that of the 103 (103/9908 tested) murine stemness homolog modules and the 22 (22/9081 tested) human stemness modules, 5 were shared between the two species, whereas the expected number of shared homolog modules was 0.17. While not many, the conserved mammalian stemness modules represent some functions crucial to stem cell function, such as cell-niche communication, self-renewal-related signaling, and chromatin remodeling. The modules that fall in the conserved category are examined in more detail in Section 7.4.

I also defined differentiation modules and identified only four differentiation homolog families and no differentiation functional modules. The differentiation modules are summarized in Table 7.3.
Table 7.1: Summary of the classification and distribution of all recurrently upregulated gene modules in the human stem cell compendium. The stemness modules (AFA and OFA pattern classes) identified by the SMA method are shown in bold. Similarly to the mouse stemness results, there is at least a 4-fold overrepresentation of AFA modules over OFA families in both the homolog and functional subsets. No significant constitutive module sets or gene sets were identified in this analysis.

### 7.3 Human stemness modules

#### 7.3.1 Angiogenesis: FGFR/FLT/PDGFR family

One of the highest scoring human stemness modules was an angiogenesis-related family of receptor tyrosine kinases. This family included the four members of the fibroblast growth factor (FGF) receptor subfamily, the two members of the platelet-derived growth factor (PDGF) receptor subfamily, the three members of the vascular endothelial growth factor (VEGF) receptor subfamily and several other receptor proteins, such as the stem cell factor (kit). These receptors have different affinities to the individual growth factors associated with each family and the difference in affinities provides the tissue and condition specificities of the growth factor receptors. As they are primarily involved with angiogenesis, many of the receptors have been implicated in cancer and differentiation regulation [31]. PDGFRα has been implicated as a marker of
<table>
<thead>
<tr>
<th>Module</th>
<th>Type</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notch; Laminin; Collagen</td>
<td>HAVA</td>
<td>2.94</td>
<td>0</td>
</tr>
<tr>
<td>Integrin alpha (ITGA)</td>
<td>HAVA</td>
<td>2.74</td>
<td>0.0015</td>
</tr>
<tr>
<td>Frizzled (Fzd)</td>
<td>HAVA</td>
<td>2.71</td>
<td>0</td>
</tr>
<tr>
<td>FGFR; FLT; PDGFR</td>
<td>HAVA</td>
<td>2.68</td>
<td>0</td>
</tr>
<tr>
<td>Pthr, Contactin</td>
<td>HAVA</td>
<td>2.66</td>
<td>0</td>
</tr>
<tr>
<td>TCF/LEF</td>
<td>HAVA</td>
<td>2.55</td>
<td>0.0203</td>
</tr>
<tr>
<td>DPP</td>
<td>HAVA</td>
<td>2.54</td>
<td>0.0076</td>
</tr>
<tr>
<td>Glypican (GPC)</td>
<td>HAVA</td>
<td>2.52</td>
<td>0</td>
</tr>
<tr>
<td>Tropomyosin</td>
<td>HAVA</td>
<td>2.50</td>
<td>0</td>
</tr>
<tr>
<td>Kinesin (KIF); Spectrin (SPTA/B)</td>
<td>HAVA</td>
<td>2.49</td>
<td>0.002</td>
</tr>
<tr>
<td>CCSPG</td>
<td>HAVA</td>
<td>2.49</td>
<td>0.0076</td>
</tr>
<tr>
<td>Chd; Smarc</td>
<td>HAVA</td>
<td>2.47</td>
<td>0</td>
</tr>
<tr>
<td>TIMP</td>
<td>HAVA</td>
<td>2.40</td>
<td>0.0031</td>
</tr>
<tr>
<td>Melanoma antigen (MAGE)</td>
<td>HAVA</td>
<td>2.33</td>
<td>0.005</td>
</tr>
<tr>
<td>Thrombospondin (THBS)</td>
<td>HAVA</td>
<td>2.28</td>
<td>0.0307</td>
</tr>
<tr>
<td>Suppressor of cytokine sig. (SOCS)</td>
<td>HAVA</td>
<td>2.23</td>
<td>0.0127</td>
</tr>
<tr>
<td>Creatine kinase (CK)</td>
<td>HAVA</td>
<td>2.21</td>
<td>0.0086</td>
</tr>
<tr>
<td>GNL</td>
<td>HAVA</td>
<td>2.20</td>
<td>0.0272</td>
</tr>
<tr>
<td>PHLDA</td>
<td>HAVA</td>
<td>2.19</td>
<td>0.0137</td>
</tr>
<tr>
<td>GABR</td>
<td>HAVA</td>
<td>2.19</td>
<td>0.011</td>
</tr>
<tr>
<td>IGFBP</td>
<td>HAVA</td>
<td>2.26</td>
<td>0.0095</td>
</tr>
<tr>
<td>DCAM</td>
<td>HAVA</td>
<td>2.22</td>
<td>0.0006</td>
</tr>
<tr>
<td>Human PPI module 181</td>
<td>FAVA</td>
<td>2.96</td>
<td>0</td>
</tr>
<tr>
<td>Human PPI module 246</td>
<td>FAVA</td>
<td>2.74</td>
<td>0</td>
</tr>
<tr>
<td>Fibril</td>
<td>FAVA</td>
<td>2.41</td>
<td>0</td>
</tr>
<tr>
<td>Protein tyrosine kinase activator activity</td>
<td>FAVA</td>
<td>2.41</td>
<td>0</td>
</tr>
<tr>
<td>Lens development in camera-type eye</td>
<td>FAVA</td>
<td>2.41</td>
<td>0.05</td>
</tr>
<tr>
<td>Positive regulation of cell-cell adhesion</td>
<td>FAVA</td>
<td>2.30</td>
<td>0</td>
</tr>
<tr>
<td>Human PPI module 98</td>
<td>FAVA</td>
<td>2.2</td>
<td>0.0307</td>
</tr>
<tr>
<td>GINS protein complex</td>
<td>FAVA</td>
<td>2.17</td>
<td>0.0076</td>
</tr>
<tr>
<td>FN1-TGM2 complex</td>
<td>FOFA</td>
<td>2.31</td>
<td>0.0026</td>
</tr>
<tr>
<td>Human PPI module 281</td>
<td>FOFA</td>
<td>2.19</td>
<td>0.0272</td>
</tr>
</tbody>
</table>

Table 7.2: List of all human stemness evolutionary and functional gene modules. First column provides the module identifier, the second column shows the classification type, the third column shows cell diversity, and the fourth column shows the recurrence FDR. The FDRs of “0” represent FDR levels below the lowest level of detectability and are thus approximately annotated with 0. Gene modules also identified in the mouse stemness meta-analysis are marked in bold. Mammalian stemness modules are based on data from similar stem cell types, such as HSC, ESC, NSC, MSC and others. Similarity between modules is based on gene membership. Comparisons between human and mouse protein-protein interaction (PPI) modules are based in human gene space, but on the data from the respective species.
neural stem cells [81], VEGFR1 is expressed in hematopoietic stem cell-enriched populations [140], while kit is one of the markers used for segregation of HSC-enriched and other populations.

### 7.3.2 Heparan sulfate proteoglycans (HSPGs): Glypican family

The heparan sulfate proteoglycan (HSPG) protein family consists of cell surface associated proteins, which can have one of several different functional cores – syndecan, glypican and perlecan. Glypican was among the highest stemness modules identified by the human stemness meta-analysis. The name of the protein family derives from the heparan sulfate side chains, associated with the core protein [39]. Interestingly, the role of glypicans is related to the receptor tyrosine kinases discussed in the previous paragraph. FGF ligands often bind to the heparan sulfate chain of these proteins before association with the FGFR receptor, which implicates the HSPG proteins as co-receptors of the various growth factor receptors [39].

<table>
<thead>
<tr>
<th>Module</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SERPIN</td>
<td>2.43</td>
<td>0</td>
</tr>
<tr>
<td>ARL</td>
<td>2.29</td>
<td>0.0241</td>
</tr>
<tr>
<td>ITGA (Itga1;Itga2;ItgaD/E/L/M/X; Itga10)</td>
<td>2.19</td>
<td>0.05</td>
</tr>
<tr>
<td>CHST</td>
<td>2.19</td>
<td>0.0243</td>
</tr>
</tbody>
</table>

Table 7.3: List of all human differentiated evolutionary and functional gene modules. First column provides the module identifier, the second column shows the classification type, the third column shows cell diversity, and the fourth column shows the recurrence FDR.
7.4 Mammalian stemness modules: Notch, TCF/LEF, Frizzled, Integrin and Chd families

7.4.1 Cell adhesion and communication: Integrin alpha family

The integrin α family, which includes the only stemness gene (Itgo6) identified in the founder studies discussed in Section 4.1.1, was consistently among the top stemness candidate families in both mouse and human. The α subfamily consists of eighteen different units, while the β subfamily includes nine different subunits [60]. The α and β subunits interact and form many different heterodimers, which represent the final functional receptors. The stemness homolog families, however, consisted of only nine of the α family members. The other nine, which included integrin αL, integrin αM, and integrin αX, were identified as a human differentiation family.

The separation of the α family into two independent homolog modules in both mouse and human reflects the somewhat different possible evolution of the α subfamilies. Integrin αL, integrin αM, and integrin αX represent members of the integrin α family that are strongly expressed in immune cells and follow more closely the evolution of the integrin β family, rather than the evolution of the rest of members of the α family [75]. This observation may explain to a certain extent why both the entire integrin β family in mouse and the integrin αL/αM/αX subfamily in human have similar expression fates to each other, completely opposite of the expression fate of the rest of the α family.

The presence of this family among the stemness modules is not entirely surprising. This family represents one of the core gene modules for communication of the cell
with its environment, so some family members are essential to the survival of the cell. For example, deletion of integrin-β1 is embryonic lethal in mice [137]. The members of the integrin family have been linked to both normal and cancer stem cell biology, as well as metastatic cancer [137].

The stemness analysis showed that in agreement with the founder studies, hematopoietic, neural, embryonic, retinal stem cells all expressed integrin α family members, but other stem cell types, such as intestinal, gastric, and trophoblast stem cells made a wide use of the family as well. Integrin α6 still remains one of the central stem cell-related proteins in this family – integrin α6 is a breast stem cell marker [154] and the integrin α6/34 is thought to adhere cells to the breast basal membrane, which represents the stem cell niche [137].

7.4.2 Wnt pathway: Tcf/LEF and Frizzled families

Consistent with the role of Wnt signaling in self-renewal, two of the five mammalian stemness modules – TCF/LEF and Frizzled – were associated with the Wnt pathway, introduced in Section 2.1.3. The stemness modules included the “beginning” and “end” of the signaling pathway, as Frizzled is the receptor on the surface of the cell that acquires the Wnt signal and begins the signal transduction process, while the TCF/LEF is the transcription factor family ultimately activated by the accumulation of β-catenin.

The stemness Tcf/LEF mouse subfamily included the Lef1, Tcf3, Tcf7, and Tcf7l2 genes, while the human family consisted of Tcf3, Tcf4, and Tcf12 in human.
Tcf3 could play a role in the regulation of the pluripotent state [172], Tcf4 may be involved in intestinal and hematopoietic stem cells maintenance [185], while Tcf12 has been implicated in neural stem and progenitor cell expansion [182].

Interestingly, I did not observe the actual Wnt ligands among the stemness modules, but since these molecules may be extensively regulated at the post-transcriptional level, their absence is not entirely surprising.

7.4.3 Chromatin: Chd/Smarca family

The Snf2 superfamily of ATP-dependent helicases plays a central role in chromatin remodeling. The proteins in this family function as part of a protein complex that has the ability to actively modify histone tails to modulate chromatin state. The superfamily has three broad subfamilies: CHD, SWI/SNF-related, and ISWI. The subfamilies are characterized by the different domains that allow them to interact with the histone tail residues – CHD proteins contain chromodomains, SWI/SNF-related proteins contain bromodomains, while ISWI proteins contain SANT domains [109]. The mouse and human stemness families consisted of two of the three subfamilies of the Snf2 superfamily: CHD and SWI/SNF-related.

The chromodomain family of proteins (Chd) consists of many chromatin remodeling enzymes. These proteins are mostly associated with marks of active transcription and are believed to maintain chromatin in an open state [58]. One of the most important stem cell-related members of this family is Chd1, which recognizes di- and tri-methylation on H3K4 (the lysine 5 residue on histone 3) residues and was recently
shown to regulate the self-renewal and the maintenance of an open chromatin state in embryonic stem cells [58]. Chd1, along with Chd3 and Chd4 showed the highest upregulation patterns in the stemness analysis, though Chd5 and Chd9 were also upregulated in a couple of different stem cell types.

The SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily A (Smarca) family was the other subfamily of proteins identified in the same stemness module. This subfamily also encodes genes with chromatin remodeling abilities, which associate with the SWI/SNF complex to regulate open chromatin. All members of the Smarca family were actively upregulated in most stem cell types in the stemness meta-analysis.

7.5 Summary

This chapter addressed the second central question of this dissertation: are stemness mechanisms conserved between mouse and human stem cells? Section 7.1 reviewed the definition of recurrently upregulated human modules, compared them to mouse recurrent modules and presented explanations for the higher heterogeneity of the human stem cell data. Section 7.2 introduced the classification of the recurrent modules into different pattern types, including the stemness modules. Section 7.3 presented a couple of the human stemness families, while Section 7.4 reviewed the main conserved mammalian stemness modules.
Chapter 8

Applications of stemness mechanisms to stem cell and cancer classification

The purpose of this chapter is to address the last question of this dissertation: can we predict the state of differentiation of a cell based on its gene expression signature? Could stem cell state be recognized and identified at the expression level? In this chapter, I directly use the stemness index defined in earlier chapters and apply it to data from various sources. Section 8.1 provides a brief overview of the test data selection. Each of the subsequent sections summarizes the results from the application of the stemness index score to a different data source – normal stem cells (Section 8.2), side populations (Section 8.3), cancer stem cells (Section 8.4), and metastatic populations (Section 8.5).
8.1 Motivation and overview

Many of the stemness modules identified by the stemness meta-analyses are of biological interest, as individual gene module members have already been implicated in various regulatory roles in self-renewal and differentiation. However, because the stemness and differentiation modules were also found to be predictive of stem cell state in a cross-validation experiment, it was natural to extend the application of predictiveness to data from new experiments, not previously included in the stem cell compendia.

I tested several heterogeneous types of data all potentially relevant to stem cell biology, as discussed previously in Chapter 2:

- Expression signatures from stem-cell-like populations
- Expression signatures from side populations
- Expression signatures from cancer stem cells
- Expression signatures from metastatic cancer populations

In the next few sections of this chapter, I present and analyze the results of the application of the stemness index classification to these new data. It should be mentioned that testing is limited to the data available in the literature and thus gathering a sufficient number of experiments in a single data type is not straightforward, as both data quality and availability are an issue. For example, metastatic cancer models in mouse that are not xenograft-based are relatively rare, as mice often die of their
primary tumor cancers before they can develop metastatic growths. The distribution of
the number of populations tested in each category is available in Table 8.1.

### 8.2 Stem cell-like populations

The most direct test of the predictive value of the stemness modules is the
performance evaluation of the stemness index score on entirely new stem cell and dif-
ferentiated populations. The new data can be particularly interesting, if they include
experiments from stem cell types that were never represented in the original stem cell
compendium. Positive results would suggest that the self-renewal signature captured
by the stem cell compendium generalizes well.

For this test, I identified 15 populations from the literature, which included
putative lung, muscle, and prostate stem cells, along with differentiated populations
from these systems. Other differentiated populations came from directed differentiation
of iPS cells into dendrite and macrophage cells. The results of the analysis of the 15

<table>
<thead>
<tr>
<th>Data type</th>
<th>Number of populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal stem/differentiated cells</td>
<td>15</td>
</tr>
<tr>
<td>Side/non-side populations</td>
<td>8</td>
</tr>
<tr>
<td>Cancer stem cell/cancer differentiated populations</td>
<td>12</td>
</tr>
<tr>
<td>Metastatic/non-metastatic populations</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total # of tested populations</strong></td>
<td><strong>39</strong></td>
</tr>
</tbody>
</table>

Table 8.1: Distribution of the sources of mouse test populations used to predict self-
renewal capacity and perform stemness classification. The first column shows each
tested data type. The second column shows the number of tested populations of each
type. The term “cancer differentiated” populations is used to refer to primary cancer
populations that are non-stem-cell-like. The total number of tested populations is shown
in bold.
populations are shown in Figures 8.1 and 8.2.

Figure 8.1 shows the comparison of three independent prostate cell signatures based on their stemness index scores. Prostate stem cell research has been an active area of research with particular relevance to prostate cancer. One study identified Sca-1 as a potential marker gene of differentiation and sorted four different populations - a primitive (embryonic) prostate stem cell population, a putative adult prostate stem cell population (Sca-1hi), a putative adult prostate progenitor population (Sca-1lo), and differentiated prostate cells Sca-1neg). To compare the potential of these populations, the authors performed gene expression comparisons of all upstream populations to the same Sca-1neg set. This experimental setup provides a unique view of self-renewal capabilities from a stemness index perspective.

The results indicated a very high stemness score for the most primitive cell populations, intermediate stemness scores for the adult prostate stem cell population and very low stemness scores for the progenitor-restricted populations (Figure 8.1). These stemness indices correlate well with the conclusions of the original study, which suggested that the most primitive cells undergo the fastest rate of self-renewal; a rate reduced after the transition to the adult stem cell type [17]. It would be particularly interesting to compare these scores to the scores of prostate cancer stem cell populations, which may revert back to the expression patterns of the more primitive population [17].

The performance of the other twelve populations is shown in Figure 8.2. Notably, the stem cell-like populations scored consistently higher than their differentiated counterparts, even though based on the stemness index cutoff selected in the initial
Figure 8.1: Stemness scores for a new mouse prostate stem cell experiment from the literature. The stemness index scores associated with each population are shown de-convoluted to their constitutive elements — the stemness and differentiation sub-scores. The line indicates a stemness index score of 0, which is the previously selected stem cell/differentiated cell classification cutoff.
mouse compendium cross-validation results, three of the five stem cell-like populations showed negative stemness indices (75% accuracy). All differentiated populations showed higher differentiation scores than stemness scores and were thus correctly classified. There are a few possible explanations for the negative stemness indices associated with three of the five stem cell populations.

One of the populations represented a mesenchymal stem cell that was described by the authors of the original study as most similar transcriptionally to mouse embryonic fibroblasts, even though it also shared some transcriptional similarities with neural and embryonic stem cells, as well as the hematopoietic stem cell niche [133]. These observations may explain its high differentiation score and reasonably high stemness score.

The second population consisted of basal cells from the mouse trachea thought to have both a self-renewal and multipotency abilities [146]. These cells are compared to a non-basal population, shown functionally to self-renew, although they did not necessarily represent a pure population. It is currently hard to assess if the negative stemness index is simply a false negative, based on the stringent cutoffs, or alternatively the tracheal population actually represents a mixture of several populations at different stages of differentiation.
Figure 8.2: Stemness scores for new mouse stem cell and differentiated cell experiments collected from the literature. The stemness index scores associated with each population are shown de-convoluted to their constitutive elements — the stemness and differentiation sub-scores. The line indicates a stemness index score of 0, which is the previously selected stem cell/differentiated cell classification cutoff.
8.3 What about side populations?

Another potential data source for stemness index application includes side populations — phenotypically defined cells based on their ability to efflux Hoechst 33342 dye, as described in Section 2.3. These populations are thought to be enriched for stem cells and have been used as means of stem cell discovery in many organs and systems.

I tested gene expression signatures from four different experiments, corresponding to eight differentially expressed sets from either side or non-side populations. It should be mentioned here that one of the side population signatures actually represented the consensus set of differentially upregulated genes from four individual side populations of bone marrow, mesenchymal, germinal and muscle origin [145]. This SP signature scored highly ($SI^1 = 0.503$) and the result is further confirmation of the validity of the stemness index approach.

Of the eight tested populations in this category, seven were correctly classified and one was potentially misclassified (87.5% accuracy). Again, the putatively misclassified population was a stem cell-enriched population — a small intestine side population thought to contain the true stem cell. It is difficult to assess the correctness of this assignment, as the population is not very well understood. In fact, while the authors of the study performed experiments to show that this stem cell-enriched population localized to the appropriate stem cell niche — the intestinal basal crypt, they did not perform any functional self-renewal assays to confirm the extent of self-renewal potential. Thus, it is possible this result is not actually a false negative, but rather a very heterogeneous

---

$^1$The SI notation refers to the stemness index score defined in Chapter 5.
mixture of stem, progenitor and even more mature cells [67], so it will be of interest to follow the future work on the purification of this population.

### 8.4 Are normal stemness mechanisms conserved in cancer stem cells?

As discussed in Chapter 2, there is already some evidence that adult normal and cancer stem cells may share molecular mechanisms, as evidenced by the use of the
same functional pathways, such as the Wnt and Notch pathways, or as indicated by the expression of the same cell surface marker genes, such as in breast or the blood system [41]. As the stemness index indirectly captures these core common processes associated with the stem cell state, it can uniquely measure the level of this conservation. I tested twelve different populations from six studies, each associated with a cancer stem cell signature and a differentiated cancer cell (non-stem-cell-like) signature.

Nine of the twelve populations were correctly classified (75% accuracy) — four of the six cancer stem cell signatures showed high stemness scores and low differentiation scores, while five of the six differentiated cancer cell signatures showed the reverse trend (Figure 8.4). The differentiated cancer cell signature generally corresponded to mature cancer cells.

Two of the misclassified (one cancer stem cell and one non-cancer stem cell) signatures in this set deserve special attention, since they highlight a point of particular relevance to the context in which the stemness index can be applied. This pair of signatures was derived from a single experiment, in which a GMP progenitor population has been transformed into a leukemic L-GMP population [101]. To aid the understanding of this point, Figure 8.5 shows a modified hematopoietic differentiation tree with the L-GMP population. The cancer stem cell signature consists of the genes commonly upregulated between the L-GMP and HSC populations, as compared to the CMP, GMP and MEP progenitor populations, while the non-stem-cell-like (differentiated) cancer cell signature contains the exact reverse set — the genes upregulated in the CMP, GMP and MEP populations. The stemness index results suggested a low differentiation score
Figure 8.4: Stemness scores for cancer stem cell populations and differentiated (non-stem-cell-like) cancer cell populations. The stemness index scores associated with each population are shown de-convoluted to their constitutive elements — the stemness and differentiation sub-scores. The line indicates a stemness index score of 0, which is the previously selected stem cell/differentiated cell classification cutoff. Cancer stem cell populations generally show very high stemness scores and low differentiation scores, suggestive of a self-renewal signature. The few misclassified populations are discussed further in the text.
Figure 8.5: The modified hematopoietic differentiation hierarchy tree shows the progressive restriction of differentiation potential from the true stem cell (LT-HSC) to more restricted progenitor cells. A leukemic stem-cell-like cell (L-GMP) could be derived from a myeloid progenitor cell, called a granulocyte-monocyte progenitor (GMP) [101].

for the non-stem-cell-like cancer signature and a high differentiation score for the cancer stem cell signature (Figure 8.4). Since the stemness index relies on differential expression information, it may be that because the subtractive populations are themselves relatively close to the stem cell in the differentiation hierarchy, the self-renewal program is harder to identify.

These results also indicated that the stemness index scoring is not appropriate for application to experiments that identify signatures from populations at equal stages of differentiation, such as a normal and a cancer mammary stem cell.
8.5 Stemness in metastasis

Another application that is relevant to stem cell biology is the assessment of the extent of self-renewal and stemness associated with metastatic cancer populations. In Chapter 2, I discussed some of the similarities between metastatic cancer cells, normal stem cells, and cancer stem cells. Specifically, these cells may all share the need for extensive self-renewal — an important feature for the successful invasion and establishment of a cancer cell in a new environment. Unfortunately, as metastatic cancer gene expression data for mouse is rarely available, I could only test four populations — two metastatic cancer populations and two primary cancer cell populations.

The results entirely concurred with the hypothesis that metastatic populations exhibit the molecular properties of stem cells, as all four populations were correctly classified based on this hypothesis (100% accuracy). The difference in stemness indices between the metastatic and non-metastatic populations was remarkably high for each set of experiments.

8.6 Summary

This chapter addressed the last question of this dissertation: can we use gene expression data to predict stem cells? I used the stemness index defined in earlier chapters and applied it to data from various sources. Section 8.1 gave a brief overview of the motivation and the data selection. Section 8.2 described the application of the stemness index score to normal stem cells. Section 8.3 tested the stemness signature in side
Figure 8.6: Stemness scores for metastatic and non-metastatic cancer populations. The stemness index scores associated with each population are shown de-convoluted to their constitutive elements — the stemness and differentiation sub-scores. The line indicates a stemness index score of 0, which is the previously selected stem cell/differentiated cell classification cutoff. Metastatic populations both show very high stemness scores and low differentiation scores, suggestive of a self-renewal signature.
populations, Section 8.4 measured the stemness index of cancer stem cell populations, and the final section, Section 8.5, evaluated the stemness of metastatic populations.
Chapter 9

Conclusion

In this final chapter of the dissertation, I summarize my findings along with some concluding thoughts and speculations. I also present my thoughts and ideas on possible future directions. Section 9.1 contains a general discussion of my findings, while Section 9.2 presents the future directions.

9.1 Discussion

This dissertation assessed computationally the validity of three independent hypotheses:

1. Functional redundancy and tissue-specific expression mask the common stem cell mechanisms.

2. Stemness mechanisms are conserved between mouse and human stem cells.

3. We can accurately predict the differentiation state of a cell based on its gene
expression signature.

To assess functional redundancy in stem cells, I developed a new methodology to test for global reproducible expression of gene modules across multiple conditions. The Stemness Meta-Analysis (SMA) method used meta-analysis techniques to identify recurrently upregulated modules across many stem cell experiments. It subsequently used other techniques to narrow the candidate stemness module list to only gene sets that were upregulated in most stem cell types and were specific to stem cells as opposed to differentiated cells. I identified 103 murine stemness modules of evolutionarily related homologous genes with reproducible, statistically significant and stem cell-specific upregulation in many mouse stem cell types. The results indicated that if we do account for functional redundancy and tissue-specific expression, previously undiscovered stemness mechanisms emerged from the mouse stem cell data.

To address the conservation of stem cell mechanisms between mouse and human cells, I also applied the stemness meta-analysis method to a human stem cell compendium. The results suggested the human data were significantly more heterogeneous than the mouse data, perhaps related to the lack of good human marker genes that may be used for isolation of pure populations. I found conservation of only five major stemness families between mouse and human: the Notch, Frizzled, Chd, TCF/LEF, and Integrin α families.

To address the predictiveness of stem cells from expression signatures, I defined a stemness index score that measures how stem cell-like a new gene expression signature is. I applied the stemness index scoring to mouse expression signatures from new stem
cell experiments, side populations, cancer stem cells, and metastatic populations. The results indicated that the mouse stemness modules, as used by the index score, can faithfully predict normal stem cells and side populations, as well as cancer stem cells and metastatic cells.

The mouse stemness modules identified in this study are the most central component of my stemness research. They provide a glimpse into “the life” of a stem cell. In earlier chapters, I mentioned that stem cells have to maintain a balance between four main states: proliferation, quiescence, apoptosis and differentiation. The stemness modules reflect this balance: I observed both proto-oncogene families and members of signaling pathways that actively promote self-renewal, also counteracted by tumor-suppressor families and signaling molecules involved in the promotion of quiescence.

By now, much evidence shows that stem cells are not solely intrinsically defined, but they also need to rely on extrinsic factors provided by the stem cell niche. Interaction with the extracellular matrix is essential for the maintenance of these cells; this necessity was reflected in the presence of adhesion molecules among the stemness modules.

Stem cells are also rare cells and not dispensable to their organ or system type; the supply of cells in each organ depends on the proper function of these rare cells. Thus, it should not be surprising that proteins associated with proper protein folding and repair, as well as DNA damage repair appear on the list of stemness modules.
9.2 Future directions

9.2.1 Application to human data

In the short term, the stemness index scoring could be applied to human data from various data sources using the human stemness modules. However, because of the small number of differentiation modules (only four) and the relatively small number of stemness modules, the human feature set may not be sufficiently informative to successfully distinguish between human stem-cell-like and non-stem-cell-like signatures. However, since many previous studies from the cancer field indicate that the mouse cancer signatures could be predictive for human cancer data, a plausible initial step could be to map expression signatures from new human experiments directly to mouse and use the mouse stemness modules to predict stemness.

9.2.2 Application to alternative splicing and miRNA data

One of the advantages of the stemness meta-analysis method is that it is a general purpose method and can be applied to any input data type to test different hypotheses. If we restrict the input profiling data to only single lineage experiments, the method could be directly used to identify lineage-specific families – markers of individual stem cell types. Such an application would resemble the efforts of the research group behind the Plurinet [117].

In addition, regulation of cell type specificity is achieved not only through homologous proteins – alternative splicing and miRNA regulation may play a significant
role as well. miRNA target prediction has been an active field of research for the last several years, but the number of miRNA targets predicted for every miRNA is usually very large. We could use the meta-analysis method structure to search for miRNA targets commonly upregulated across many differentiated cells, which may suggest miRNAs highly active in stem cells.

Another possible application, given solid expression test data is to define families of alternatively spliced isoforms, instead of homolog protein modules. Many proteins are known to use alternatively spliced isoforms to induce tissue specificity to various cell types. This application, however, may be more long-term as alternative splicing data for various stem cell types is not available yet.

### 9.2.3 Addition of niche data

Stem cell niches play a central role in the maintenance of stem cells. Even though niches are better understood at present than they were a decade ago, very few profiling experiments exist that examine the expression of their cells. Given data from various stem cell niches, it would be very interesting to compare the stemness and potential “niche-ness” signatures, as I expect that the cross-talk between stem and niche cells would yield many common factors.

### 9.2.4 Methodological improvements

One aspect that the current SMA method does not take into account is the possible specialization of different genes within a module to cells at various stages of
differentiation. Because of that, it is possible to observe stemness families that have member genes with a non-negligible upregulation in differentiated cells. We could define a score that measures the specificity level of each gene in a given stemness module to stem cells or differentiated cells.

A more important improvement over the current method would be to use phylogenetically informed homolog modules. While the single e-value alignment cutoff I used to define homolog families worked well for many gene modules, some homolog group assignments could be improved. A paralog/homolog family-specific cutoff that accounts for the evolution of the family would be certainly more informative, though non-trivial to define.

9.2.5 Possible biological experiments

More than a few proteins from the set of stemness modules have already been implicated as self-renewal genes or as potential master regulators of differentiation, and many more could still be tested. Transcription factor stemness modules could be used to test for new reprogramming candidates.

RNAi-mediated silencing of individual stemness family members could be used to test directly their role in differentiation, in a manner similar to the one used by Suzuki et al. [169], where RNAi was used to silence Myb. The effects of Myb silencing resembled very closely the effects observed after treatment with a differentiation-inducing drug [169].

Since many of these proteins have been also implicated in metastatic can-
cer, Matrigel invasiveness assays could be used to test the metastatic abilities of cells that have had individual stemness family members knocked out, or otherwise silenced. Invasiveness abilities could also be tested through stemness gene (module) activation (knock-in) experiments in normal cancer cell lines, or alternatively through drug inhibition of genes from stemness modules in metastatic cancer lines.
Appendix A

Definitions of key terms

**Adenocarcinoma**  Cancer of the protective epithelial cells that line internal organs.

**Apoptosis**  Cellular death associated with normal cell turnover. Cancer cells generally lose their ability to regulate this process.

**Carcinoma**  Malignant cancer of the epithelial cells.

**Cancer stem cells**  Rare cancer cells that have the ability to self-renew and give rise to differentiated cancer cells. These cells are thought to be at the top of the differentiation tree in the cancer cell hierarchy, just like normal stem cells are at the top of the normal cell differentiation hierarchy.

**Cell-type diversity**  An entropy-based score that measures how evenly distributed the upregulation of a gene module is across different cell types.

**Differentiation**  The process through which mature cells arise from less differentiated stem and progenitor cells.
**Epithelial-mesenchymal transition (EMT)** A morphological and phenotypic transition from an epithelial state (adherent, structured, not migratory) to a mesenchymal state (less structured and more migratory). This transition is essential for a successful invasion of a new tissue and is considered to be one of the hallmarks of metastatic cancer.

**FACS** Fluorescence-activated cell-sorting flow cytometry is a widely used technique in the stem cell field. FACS is used to isolate pure populations based on their cell surface expression signature pattern.

**Gene-usage diversity** An entropy-based score that measures how evenly distributed the upregulation of the genes in a gene module is.

**Induced pluripotent stem (iPS) cells** Cells derived from fully differentiated cells, which upon induction with one or more transcription factors, such as Myc or Sox2, can transform into pluripotent cells, bearing all the hallmarks of embryonic stem cells.

**Inner cell mass** The cells that give rise to all three germ layers – ectoderm, mesoderm and endoderm – of the embryo, but not to the extra-embryonic tissue layer. Embryonic stem cells are derived from the inner cell mass.

**Inverse-variance weighting** A technique used in standard meta-analysis, where the contribution of each study to an overall combined effect is weighted by the inverse of the variance associated with the measured effect in the study.
Meta-analysis A field of statistics that focuses on the inference of a measurable combined effect through the integration of the individual effects measured in many different studies.

Metastatic tumor Tumors that have evolved from a primary tumor, but have invaded a more distant and unrelated to the primary tumor tissue.

Monoclonal Arising from a single common ancestral cell.

Multipotent cells Cells that have the ability to give rise to only the cells within a particular organ or system type. This term is most often associated with adult stem cells, such as hematopoietic or neural stem cells.

Pluripotent cells Cells that have the ability to give rise to any cell with an organism with the exception of the extra-embryonic tissue layer that forms the placenta. The cells that are included in this category can make all three embryonic developmental layers: ectoderm, mesoderm and endoderm.

Primary tumors Tumors that develop at the site of origin of the ancestral cell mutation

Proliferation The process of active self-renewal of normal and cancer stem cells.

Quiescence An inactive non-proliferative stem cell state.

Recurrence Reproducible upregulation of a module or gene across many studies or experiments.
**Side population** A stem cell-enriched population that is characterized by its ability to efflux drugs, toxins, or dyes using the ABC family of transporter proteins.

**Squamous cell carcinoma** Cancer of the secretory cells whose purpose is to release secretions that protect the epithelial linings of internal organs.

**Stem cells** Functionally defined undifferentiated cells that are characterized by their ability to self-renew and give rise to many mature cell types.

**Trophectoderm** The cells that form the extra-embryonic tissues and give rise to the placenta.

**Tumor** An abnormal growth of cells that lose the ability to undergo controlled proliferation.

**Tumor heterogeneity** Subtype, individual organism, or cellular disparity between the cells in a single tumor growth.
Appendix B

Tables of stemness modules

<table>
<thead>
<tr>
<th>Homolog AFA module name</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notch; Laminin; Collagen</td>
<td>3.37</td>
<td>0.006</td>
</tr>
<tr>
<td>Kif2</td>
<td>3.20</td>
<td>0.011</td>
</tr>
<tr>
<td>Myh</td>
<td>3.15</td>
<td>0.021</td>
</tr>
<tr>
<td>Tubulin</td>
<td>3.11</td>
<td>0</td>
</tr>
<tr>
<td>Prps</td>
<td>3.10</td>
<td>0</td>
</tr>
<tr>
<td>Hdgf</td>
<td>3.08</td>
<td>0.014</td>
</tr>
<tr>
<td>Chd/Smarc</td>
<td>3.06</td>
<td>0</td>
</tr>
<tr>
<td>Ppi</td>
<td>3.02</td>
<td>0.009</td>
</tr>
<tr>
<td>Ephrin</td>
<td>3.02</td>
<td>0.046</td>
</tr>
<tr>
<td>Frizzled</td>
<td>3.01</td>
<td>0.008</td>
</tr>
<tr>
<td>Ncl/Supt</td>
<td>3.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Sfrrs</td>
<td>2.98</td>
<td>0</td>
</tr>
<tr>
<td>Grb</td>
<td>2.96</td>
<td>0.01</td>
</tr>
<tr>
<td>Nap1</td>
<td>2.93</td>
<td>0.011</td>
</tr>
<tr>
<td>Rcc/Herc</td>
<td>2.93</td>
<td>0.036</td>
</tr>
<tr>
<td>Wdr</td>
<td>2.93</td>
<td>0</td>
</tr>
<tr>
<td>Nme</td>
<td>2.92</td>
<td>0.039</td>
</tr>
<tr>
<td>Mapk</td>
<td>2.91</td>
<td>0.001</td>
</tr>
<tr>
<td>Sfrrs4–6</td>
<td>2.88</td>
<td>0.005</td>
</tr>
<tr>
<td>Myb</td>
<td>2.87</td>
<td>0</td>
</tr>
<tr>
<td>Plk</td>
<td>2.87</td>
<td>0.007</td>
</tr>
<tr>
<td>Hsp (heat shock proteins)</td>
<td>2.83</td>
<td>0.001</td>
</tr>
<tr>
<td>Abr/Bcr</td>
<td>2.82</td>
<td>0</td>
</tr>
<tr>
<td>Mcm</td>
<td>2.81</td>
<td>0</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Homolog</th>
<th>AFA module name</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ccnb</td>
<td></td>
<td>2.80</td>
<td>0.0002</td>
</tr>
<tr>
<td>Rcn</td>
<td></td>
<td>2.80</td>
<td>0.007</td>
</tr>
<tr>
<td>Terf</td>
<td></td>
<td>2.79</td>
<td>0.021</td>
</tr>
<tr>
<td>Myc</td>
<td></td>
<td>2.79</td>
<td>0</td>
</tr>
<tr>
<td>Aurora</td>
<td></td>
<td>2.77</td>
<td>0</td>
</tr>
<tr>
<td>Dusp1–4</td>
<td></td>
<td>2.77</td>
<td>0.007</td>
</tr>
<tr>
<td>Fus/Nola</td>
<td></td>
<td>2.77</td>
<td>0.015</td>
</tr>
<tr>
<td>Id (inhibitor of differentiation)</td>
<td>2.76</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td>p53 (p53, p63, p73)</td>
<td>2.75</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>Kpna</td>
<td></td>
<td>2.74</td>
<td>0.027</td>
</tr>
<tr>
<td>Itga (Integrin alpha)</td>
<td>2.74</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Pbx</td>
<td></td>
<td>2.73</td>
<td>0.003</td>
</tr>
<tr>
<td>F2r</td>
<td></td>
<td>2.73</td>
<td>0.006</td>
</tr>
<tr>
<td>Smo</td>
<td></td>
<td>2.72</td>
<td>0.005</td>
</tr>
<tr>
<td>Camk</td>
<td></td>
<td>2.72</td>
<td>0.049</td>
</tr>
<tr>
<td>Ldha</td>
<td></td>
<td>2.71</td>
<td>0.009</td>
</tr>
<tr>
<td>Cdkn1a–c (p21,p27,p57)</td>
<td>2.71</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Ccne</td>
<td></td>
<td>2.70</td>
<td>0.0001</td>
</tr>
<tr>
<td>Rbp</td>
<td></td>
<td>2.70</td>
<td>0</td>
</tr>
<tr>
<td>Cct (chaperonin)</td>
<td>2.69</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sfrp</td>
<td></td>
<td>2.69</td>
<td>0.0005</td>
</tr>
<tr>
<td>Tie</td>
<td></td>
<td>2.69</td>
<td>0.006</td>
</tr>
<tr>
<td>Brca1/Mcph1</td>
<td>2.69</td>
<td>.009</td>
<td></td>
</tr>
<tr>
<td>Aesm/Acsf</td>
<td>2.68</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>Spred</td>
<td></td>
<td>2.68</td>
<td>0.033</td>
</tr>
<tr>
<td>Lypla</td>
<td></td>
<td>2.68</td>
<td>0.011</td>
</tr>
<tr>
<td>Fen1/Gen1</td>
<td>2.67</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Ccn</td>
<td></td>
<td>2.66</td>
<td>0.007</td>
</tr>
<tr>
<td>Slec2a</td>
<td></td>
<td>2.65</td>
<td>0.009</td>
</tr>
<tr>
<td>Pkp</td>
<td></td>
<td>2.65</td>
<td>0.0005</td>
</tr>
<tr>
<td>Meis</td>
<td></td>
<td>2.63</td>
<td>0.0003</td>
</tr>
<tr>
<td>Hnrnpa</td>
<td></td>
<td>2.63</td>
<td>0.01</td>
</tr>
<tr>
<td>Hnrnpd</td>
<td></td>
<td>2.63</td>
<td>0.0002</td>
</tr>
<tr>
<td>Lim</td>
<td></td>
<td>2.61</td>
<td>0.019</td>
</tr>
<tr>
<td>Hmgb (high mobility)</td>
<td>2.60</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Khdrbs</td>
<td></td>
<td>2.59</td>
<td>0.033</td>
</tr>
<tr>
<td>Ctbp</td>
<td></td>
<td>2.59</td>
<td>0.007</td>
</tr>
<tr>
<td>Mkrm</td>
<td></td>
<td>2.58</td>
<td>0.026</td>
</tr>
<tr>
<td>Bub1/1b</td>
<td></td>
<td>2.58</td>
<td>0.002</td>
</tr>
<tr>
<td>Rnfl25/138/166</td>
<td>2.57</td>
<td>0.007</td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
### Table B.1 – continued from previous page

<table>
<thead>
<tr>
<th>Homolog AFA module name</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eva (eyeless)</td>
<td>2.57</td>
<td>0.005</td>
</tr>
<tr>
<td>Ipo (importin)</td>
<td>2.57</td>
<td>0</td>
</tr>
<tr>
<td>Acot</td>
<td>2.56</td>
<td>0.033</td>
</tr>
<tr>
<td>Ctnna</td>
<td>2.56</td>
<td>0.007</td>
</tr>
<tr>
<td>Cks</td>
<td>2.56</td>
<td>0.0003</td>
</tr>
<tr>
<td>Gnb</td>
<td>2.56</td>
<td>0.011</td>
</tr>
<tr>
<td>Msh</td>
<td>2.56</td>
<td>0.0035</td>
</tr>
<tr>
<td>Lancl</td>
<td>2.56</td>
<td>0.026</td>
</tr>
<tr>
<td>Lsp1/Cdc113</td>
<td>2.55</td>
<td>0.004</td>
</tr>
<tr>
<td>Ube2/Birc</td>
<td>2.55</td>
<td>0.023</td>
</tr>
<tr>
<td>Ssbp</td>
<td>2.53</td>
<td>0.012</td>
</tr>
<tr>
<td>Kif20/Kif23</td>
<td>2.51</td>
<td>.001</td>
</tr>
<tr>
<td>Tcf/LEF</td>
<td>2.51</td>
<td>0.0009</td>
</tr>
<tr>
<td>Ilf</td>
<td>2.50</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Table B.1: List of mouse stemness homolog AFA modules.

### Functional AFA module name

<table>
<thead>
<tr>
<th>Functional AFA module name</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Go:liver development</td>
<td>3.07</td>
<td>0</td>
</tr>
<tr>
<td>Go:regulation of TGFβ pathway</td>
<td>3.03</td>
<td>0.004</td>
</tr>
<tr>
<td>Mouse PPI Interactions Module 11-4-5</td>
<td>3.03</td>
<td>0.001</td>
</tr>
<tr>
<td>Go:blastocyst growth</td>
<td>3.02</td>
<td>0.022</td>
</tr>
<tr>
<td>Go:establishment of planar polarity</td>
<td>2.95</td>
<td>0.011</td>
</tr>
<tr>
<td>Human PPI Interactions Module147-11-15</td>
<td>2.93</td>
<td>0</td>
</tr>
<tr>
<td>Human PPI Interactions Module251-11-14</td>
<td>2.90</td>
<td>0</td>
</tr>
<tr>
<td>Human PPI Interactions Module3-21-54</td>
<td>2.89</td>
<td>0</td>
</tr>
<tr>
<td>Go:cartilage condensation</td>
<td>2.89</td>
<td>0.021</td>
</tr>
<tr>
<td>Go:determination of ante/post axis embryo</td>
<td>2.89</td>
<td>0.001</td>
</tr>
<tr>
<td>Human PPI Interactions Module96-14-31</td>
<td>2.88</td>
<td>0.001</td>
</tr>
<tr>
<td>Go:female pronucleus</td>
<td>2.86</td>
<td>0.008</td>
</tr>
<tr>
<td>Human PPI Interactions Module54-16-32</td>
<td>2.85</td>
<td>0</td>
</tr>
<tr>
<td>Go:RNA helicase activity</td>
<td>2.84</td>
<td>0</td>
</tr>
<tr>
<td>Go:embryonic cleavage</td>
<td>2.81</td>
<td>0.014</td>
</tr>
<tr>
<td>Human PPI Interactions Module258-10-14</td>
<td>2.80</td>
<td>0.021</td>
</tr>
<tr>
<td>Go:neuromuscular synaptic transmission</td>
<td>2.79</td>
<td>0.033</td>
</tr>
<tr>
<td>Go:morphogenesis of an epithelial sheet</td>
<td>2.79</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Functional AFA module name</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Go:heart morphogenesis</td>
<td>2.77</td>
<td>0</td>
</tr>
<tr>
<td>Human PPI Interactions Module171-4-6</td>
<td>2.74</td>
<td>0</td>
</tr>
<tr>
<td>Go:paraxial mesoderm morphogenesis</td>
<td>2.71</td>
<td>0.0335</td>
</tr>
<tr>
<td>Human PPI Interactions Module263-5-5</td>
<td>2.70</td>
<td>0.024</td>
</tr>
<tr>
<td>BC:Transcr. activation of dbpb from mRNA</td>
<td>2.69</td>
<td>0.015</td>
</tr>
<tr>
<td>Human PPI Interactions Module182-6-7</td>
<td>2.69</td>
<td>0.039</td>
</tr>
<tr>
<td>Go:centric heterochromatin</td>
<td>2.68</td>
<td>0</td>
</tr>
<tr>
<td>Go:methylation-dependent chromatin silencing</td>
<td>2.68</td>
<td>0</td>
</tr>
<tr>
<td>Human PPI Interactions Module236-5-5</td>
<td>2.67</td>
<td>0.008</td>
</tr>
<tr>
<td>Human PPI Interactions Module115-18-40</td>
<td>2.67</td>
<td>0</td>
</tr>
<tr>
<td>Bcl-xL-p53-PUMA DNA damage complex</td>
<td>2.67</td>
<td>0.0145</td>
</tr>
<tr>
<td>Go:pyrimidine base metabolic process</td>
<td>2.67</td>
<td>0.002</td>
</tr>
<tr>
<td>Go:ER-Golgi intermediate compartment</td>
<td>2.65</td>
<td>0.0238</td>
</tr>
<tr>
<td>DNA polymerase delta complex</td>
<td>2.65</td>
<td>0.0004</td>
</tr>
<tr>
<td>Go:neg. regulation of protein import into nucleus</td>
<td>2.65</td>
<td>0.005</td>
</tr>
<tr>
<td>Go:mitochondrial ribosome</td>
<td>2.64</td>
<td>0</td>
</tr>
<tr>
<td>Go:cell structure disassembly during apoptosis</td>
<td>2.63</td>
<td>0.0362</td>
</tr>
<tr>
<td>Human PPI Interactions Module172-4-5</td>
<td>2.62</td>
<td>0.0009</td>
</tr>
<tr>
<td>Go:nuclear speck</td>
<td>2.61</td>
<td>0.007</td>
</tr>
<tr>
<td>Go:cytosolic small rib. subunit sen. Eukaryota</td>
<td>2.61</td>
<td>0.0015</td>
</tr>
<tr>
<td>HumanPPI Interactions Module264-14-23</td>
<td>2.60</td>
<td>0</td>
</tr>
<tr>
<td>Go:cell migration involved in gastrulation</td>
<td>2.57</td>
<td>0.009</td>
</tr>
<tr>
<td>Go:cysteine metabolic process</td>
<td>2.56</td>
<td>0.0075</td>
</tr>
<tr>
<td>Human PPI Interactions Module232-4-4</td>
<td>2.56</td>
<td>0.045</td>
</tr>
<tr>
<td>Go:imprinting</td>
<td>2.56</td>
<td>0</td>
</tr>
<tr>
<td>Go:DS break repair via hom recombination</td>
<td>2.56</td>
<td>0</td>
</tr>
<tr>
<td>Go:nuclear lamina</td>
<td>2.53</td>
<td>0.0008</td>
</tr>
<tr>
<td>Human PPI Interactions Module228-4-6</td>
<td>2.50</td>
<td>0.045</td>
</tr>
</tbody>
</table>

Table B.2: List of mouse stemness functional AFA modules.
<table>
<thead>
<tr>
<th>Homolog OFA module name</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orc1</td>
<td>2.87</td>
<td>0</td>
</tr>
<tr>
<td>Rad51</td>
<td>2.80</td>
<td>0.001</td>
</tr>
<tr>
<td>Impdh</td>
<td>2.80</td>
<td>0.005</td>
</tr>
<tr>
<td>Camkk</td>
<td>2.80</td>
<td>0.008</td>
</tr>
<tr>
<td>Bzw</td>
<td>2.78</td>
<td>0.048</td>
</tr>
<tr>
<td>Atad/Fignl</td>
<td>2.77</td>
<td>0.009</td>
</tr>
<tr>
<td>Ckap</td>
<td>2.73</td>
<td>0</td>
</tr>
<tr>
<td>Shroom</td>
<td>2.71</td>
<td>0.001</td>
</tr>
<tr>
<td>G3bp</td>
<td>2.70</td>
<td>0.009</td>
</tr>
<tr>
<td>Cct8/Gm443</td>
<td>2.68</td>
<td>0.040</td>
</tr>
<tr>
<td>Steap</td>
<td>2.66</td>
<td>0.0238</td>
</tr>
<tr>
<td>Shmt</td>
<td>2.64</td>
<td>0</td>
</tr>
<tr>
<td>Rrm2</td>
<td>2.64</td>
<td>0.003</td>
</tr>
<tr>
<td>Top2</td>
<td>2.64</td>
<td>0.012</td>
</tr>
<tr>
<td>Pa2g4/Metap2</td>
<td>2.63</td>
<td>0.012</td>
</tr>
<tr>
<td>Set</td>
<td>2.62</td>
<td>0.009</td>
</tr>
<tr>
<td>Fbl</td>
<td>2.58</td>
<td>0.016</td>
</tr>
<tr>
<td>H2af</td>
<td>2.56</td>
<td>0.014</td>
</tr>
<tr>
<td>Csrp</td>
<td>2.56</td>
<td>0.0099</td>
</tr>
<tr>
<td>Galk</td>
<td>2.52</td>
<td>0.001</td>
</tr>
<tr>
<td>Ruvbl</td>
<td>2.50</td>
<td>0</td>
</tr>
</tbody>
</table>

Table B.3: List of mouse stemness homolog OFA modules.

<table>
<thead>
<tr>
<th>Functional OFA module name</th>
<th>Cell diversity</th>
<th>Recurrence FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Go:hydroxymethyl-formyl related transf. activity</td>
<td>2.91</td>
<td>0.005</td>
</tr>
<tr>
<td>Go:internal protein amino acid acetylation</td>
<td>2.90</td>
<td>0.0005</td>
</tr>
<tr>
<td>Rag1-Rag2-Ku70-Ku80 protein-DNA complex</td>
<td>2.79</td>
<td>0.014</td>
</tr>
<tr>
<td>Go:DNA replication synthesis of RNA primer</td>
<td>2.73</td>
<td>0.003</td>
</tr>
<tr>
<td>PU1-associated protein complex</td>
<td>2.72</td>
<td>0.011</td>
</tr>
<tr>
<td>Go:nerve growth factor receptor signaling pathway</td>
<td>2.63</td>
<td>0.005</td>
</tr>
<tr>
<td>Go:dosage compensation by inactivation of X chrom</td>
<td>2.61</td>
<td>0.005</td>
</tr>
<tr>
<td>Go:pos. regulation of gene-specific transcription</td>
<td>2.59</td>
<td>0.0008</td>
</tr>
<tr>
<td>Human PPI Interactions Module177-4-4</td>
<td>2.59</td>
<td>0.0108</td>
</tr>
<tr>
<td>Cd2ap-Fyn complex</td>
<td>2.58</td>
<td>0.008</td>
</tr>
<tr>
<td>Go:NF-kappaB import into nucleus</td>
<td>2.56</td>
<td>0.023</td>
</tr>
<tr>
<td>Go:regulation of DNA replication initiation</td>
<td>2.55</td>
<td>0.001</td>
</tr>
<tr>
<td>Go:positive regulation of exocytosis</td>
<td>2.52</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table B.4: List of mouse stemness functional OFA modules.

217
Bibliography


[69] Xi C He, Jiawang Zhang, Wei-Gang Tong, Ossama Tawfik, Jason Ross, David H Scoville, Qiang Tian, Xin Zeng, Xi He, Leanne M Wiedemann, Yuji Mishina, and Linheng Li. BMP signaling inhibits intestinal stem cell self-renewal through


[123] Kazuya Ogawa, Akira Saito, Hisanori Matsui, Hiroshi Suzuki, Satoshi Ohitsuka, Daisuke Shimosato, Yasuyuki Morishita, Tetsuro Watabe, Hitoshi Niwa, and Ko-
hei Miyazono. Activin-Nodal signaling is involved in propagation of mouse em-

[124] T Ohitsuka, M Ishibashi, G Gradwohl, S Nakanishi, F Guillemot, and R Kageyama. Hes1 and Hes5 as notch effectors in mammalian neuronal differen-

[125] J Okabe-Kado, T Kasukabe, and Y Honma. Differentiation inhibitory factor
Nm23 as a prognostic factor for acute myeloid leukemia. Leuk Lymphoma,

Mack, Mary R Avarbock, Lewis Chodosh, and Ralph L Brinster. Genes in-
volved in post-transcriptional regulation are overrepresented in stem/progenitor

and renewal of hair follicles from adult multipotent stem cells. Cell, 104(2):233–
245, Jan 2001.

[128] V E Papaioannou, M W McBurney, R L Gardner, and M J Evans. Fate of
73, Nov 1975.

[129] Ricardo Pardal, Michael F Clarke, and Sean J Morrison. Applying the principles

[130] In-Hyun Park, Rui Zhao, Jason A West, Akiko Yabuuchi, Hongguang Huo, Tan A
Ince, Paul H Lerou, M William Lensch, and George Q Daley. Reprogramming of
146, Jan 2008.

[131] In-kyung Park, Dalong Qian, Mark Kiel, Michael W Becker, Michael Pihalja,
Irving L Weissman, Sean J Morrison, and Michael F Clarke. Bmi-1 is re-
quired for maintenance of adult self-renewing haematopoietic stem cells. Nature,

[132] Helen Parkinson, Misha Kapushesky, Nikolay Kolesnikov, Gabriella Rustici,
Mohammad Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dy-
lag, Ibrahim Emam, Anna Farne, Ele Holloway, Margus Lukk, James Malone,
Roby Mani, Ekaterina Plicheva, Tim F Rayner, Faisal Rezwan, Anjan Sharma,
Eleanor Williams, Xiangqun Zheng Bradley, Tomasz Adamusiak, Marco Brandizi,
Tony Burdett, Richard Coulson, Maria Krestyaninova, Pavel Kurnosov, Eamonn
Maguire, Sudeshna Guha Neogi, Philippe Rocca-Serra, Susanna-Assunta Sansone,


233


[172] Wai-Leong Tam, Chin Yan Lim, Jianyong Han, Jinqiu Zhang, Yen-Sin Ang, Huck-Hui Ng, Henry Yang, and Bing Lim. T-cell factor 3 regulates embryonic stem cell


