The Secret of Stem Cells

Transcriptional Profiling of Human Stem Cells

Michal Golan-Mashiach
M.Sc Thesis submitted to the Feinberg Graduate School
Weizmann Institute of Science

Research conducted under the supervision of
Prof. Eytan Domany and Prof. David Givol

December 2003

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

First and foremost I am grateful for the guidance, advice and enthusiastic supervision of Prof. Eytan Domany during this work. His perseverance as a physicist in quantifying biological phenomena leaving no stone unturned is inspiring.

I owe special thanks to Prof. David Givol, a talented teacher and passionate scientist, for his endless patience and guidance from the day I first met him. His enthusiasm and joy of research were inspiring and contagious. He is truly one of my greatest teachers ever.

The work presented here was done in collaboration with several people: Dr. Jean-Eudes Dazard who was a great teacher in the ``wet-lab'' work for the skin cells, Dr. Gideon Rechavi and his lab members for the blood cells and Affymetrix experiment and Dr. Joseph Eldor-Itskovitz and his lab members for the embryonic cells.

I would like to thank all my friends from Domany's lab and others, who were like family to me during this year: Dr. Gaddy Getz, Omer Barad, Hilah Gal, Ilan & Dafna Tsafrir, Hertzberg Libi, Or Zuk, Gadi Elizur, Hila Benjamin Rodrig, Uri Einav, Yuval Tabach, Assif Yitzhaky, Liat Ein Dor, Barzuza Tamar and Gorelick Yelena.

Finally, I am forever indebted to my husband Erez for his thoughts, ideas, understanding, endless patience and encouragement when it was most required. I could have never done this without him.

Contents

Chapter 1

1 Biological Background

1.1 Definitions, Concepts, Community and Medical Interest about Stem Cells

A stem cell is a special kind of cell that has a unique capacity to renew itself and to give rise to many specialized cell types (pluripotency). Contrary to most cells of the body, such as heart cells or skin cells, which are committed to perform a specific function, a stem cell is uncommitted and remains uncommitted, until it receives a signal to develop into a specialized cell. Work in this field includes two kinds of stem cells from animals and humans: embryonic stem cells and adult stem cells, which have different functions and characteristics.

Learning about stem cells can be used for specific purposes: using the cells in cell-based therapies and in genetic engineering or gene therapy [1], screening new drugs and toxins and understanding birth defects [2]. However, human embryonic stem cells have been studied only since 1998 [3]. In order to develop such treatments, one has to first concentrate on the fundamental properties of stem cells, which include: 1) determining precisely how stem cells remain unspecialized and self renewing for many years; and 2) identifying the signals that cause stem cells to become specialized.

Stem cells are important for living organisms for many reasons. In the 3 to 5 day old embryo, called a blastocyst, a small group of about 30 cells called the inner cell mass gives rise to the billions of highly specialized cells needed to make up an adult organism. In the developing fetus, stem cells give rise to the multiple specialized cell types that make up the heart, lung, skin, and other tissues (see fig. 1.1). In some adult tissues, such as bone marrow, muscle, and brain, small populations of adult stem cells generate replacements for cells that are lost through normal wear and tear, injury, or disease [7, 8]. It has been hypothesized that stem cells may, at some point in the future, become the basis for treating diseases such as Parkinson's disease, diabetes, and heart disease by therapeutic transplantation. This may open ways for tissue damage repair in "personalized medicine". There are several approaches to study a stem cell. One can start from the phenotype aspect, which refers to all the observable characteristics of a cell (or organism); its shape (morphology); interactions with other cells and the non-cellular environment (also called the extracellular matrix); proteins that appear on the cell surface (surface markers); and the cell's behavior (e.g., secretion, contraction, synaptic transmission). Alternatively, one can study through functionality, which refers to the genetic profiles or transcriptomes of the cell. Stem cells are one of the most fascinating areas of biology today. But like many expanding fields of scientific inquiry, research on stem cells raises scientific questions as rapidly as it generates new discoveries. There are many ways in which human stem cells can be used in basic and clinical research. However, there are many technical hurdles between the promise of stem cells and the realization of these uses, which will be overcome by continued intensive stem cell research.

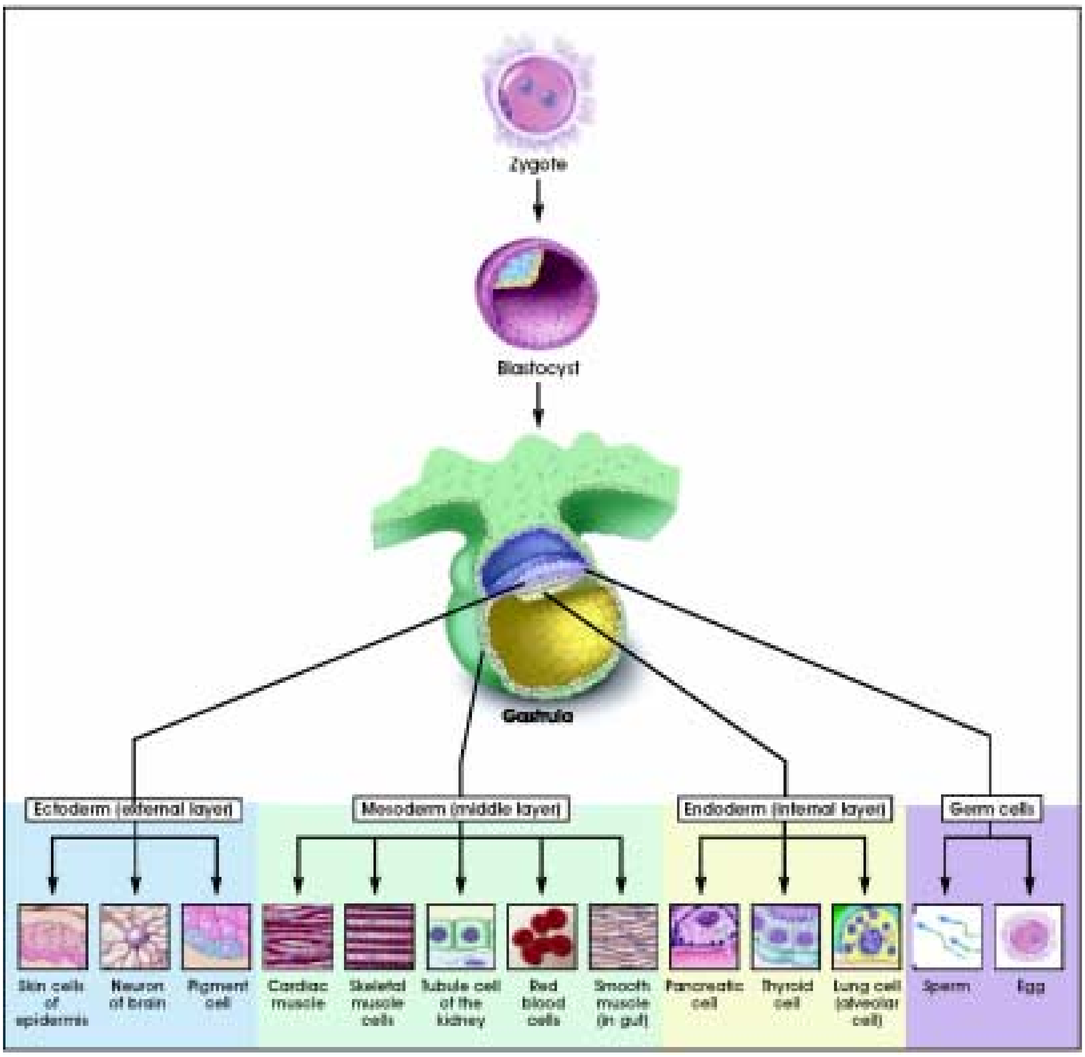


Figure 1.1 Differentiation of Human Tissues Three embryonic germ layers – mesoderm, endoderm and ectoderm are the source of all cells of the body. All the different kinds of specialized cells that make up the body are derived from one of these germ layers.

1.2 What are the unique properties of all stem cells?

Stem cells differ from other kinds of cells in the body. All stem cells — regardless of their source — have three general properties: they are capable of dividing and renewing themselves for long periods; they are unspecialized; and they can give rise to specialized cell types [9].

Stem cells are capable of dividing and renewing themselves for long periods. Unlike most mature cells e.g. muscle cells, blood cells, or nerve cells — which do not normally replicate— stem cells may replicate many times. Many repeated replication of a cell is called proliferation. A starting population of stem cells, that proliferates for many months in the laboratory, can yield millions of cells. If the resulting cells continue to be unspecialized, like the parent stem cells, the cells are said to be capable of long-term self-renewal.

The specific factors and conditions that allow stem cells to remain unspecialized are of great interest. Therefore, an important area of research is understanding the signals in a mature organism that cause a stem cell population to proliferate and remain unspecialized until the cells are needed for the normal process of replacement of dead cells e.g. skin and colon, or for repair of a specific damaged tissue. Such information is critical to be able to grow large numbers of unspecialized stem cells in the laboratory for further experimentation.

Stem cells are unspecialized. One of the fundamental properties of a stem cell is that it does not have any tissue-specific structures that allow it to perform specialized functions. A stem cell cannot work with its neighbors to pump blood through the body (like a heart muscle cell); it cannot carry molecules of oxygen through the bloodstream (like a red blood cell); and it cannot fire electrochemical signals to other cells that allow the body to move or speak (like a nerve cell). However, unspecialized stem cells can give rise to specialized cells, including heart muscle cells, blood cells, or nerve cells.

Stem cells can give rise to specialized cells. The process by which unspecialized stem cells give rise to specialized cells is called differentiation. The main questions are to understand the signals from within and from outside the cells, that trigger stem cell differentiation. The internal signals are controlled by a cell's genes. The external signals for cell differentiation include chemicals secreted by other cells, physical contact with neighboring cells, and certain molecules in the microenvironment. To date, several laboratories have demonstrated that human embryonic stem cells in vitro are pluripotent; they can produce cell types derived from three embryonic germ layers (endoderm, mesoderm and ectoderm) [4-6, 10].

Many questions about stem cell differentiation remain open. For example, are the internal and external signals for cell differentiation similar for all kinds of stem cells? Can specific sets of signals be identified that promote differentiation into specific cell types? Addressing these questions is critical because the answers may lead us to find new ways of controlling stem cell differentiation in the laboratory, thereby growing cells or tissues that can be used for specific purposes, including cell-based therapies.

1.3 The Embryonic Stem cell

Embryonic stem cells, as their name suggests, are derived from embryos. Specifically, embryonic stem cells are derived from embryos that develop from eggs that have been fertilized in vitro — in an in vitro fertilization clinic — and then donated for research purposes with informed consent of the donors [11]. The embryos from which human embryonic stem cells are derived are typically four or five days old and are a hollow microscopic ball of cells called the blastocyst. The blastocyst includes three structures: the trophoblast, which is the layer of cells that surrounds the blastocyst; the blastocoel, which is the hollow cavity inside the blastocyst; and the inner cell mass, which is a group of approximately 30 cells at one end of the blastocoel (see Fig 1.2).

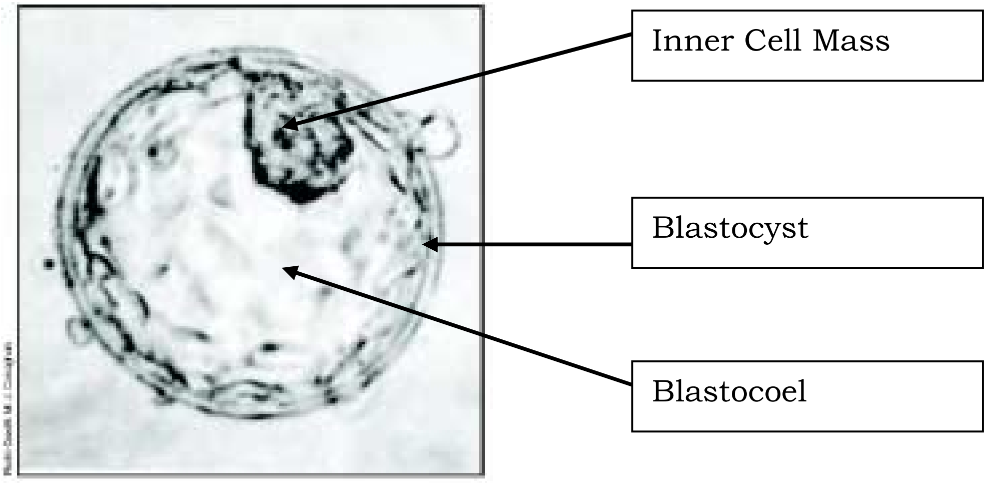


Figure 1.2 Human Blastocyst, which is the pre-implantation embryos containing ~150 cells, showing Inner Cell Mass (ICM) and trophectoderm.

In 1981, there have been reports [12] of methods for growing mouse embryonic stem cells in the laboratory; it took nearly 20 years before similar achievements could be made with human embryonic stem cells. In 1998, James Thomson and his colleagues reported methods for deriving and maintaining human embryonic stem cells from the inner cell mass of human blastocysts that were produced through in vitro fertilization and donated for research purposes [3]. At the same time, another group, led by John Gearhart, reported the derivation of cells that they identified as embryonic germ cells. The cells were cultured from primordial germ cells obtained from the gonadal ridge and mesenchymal cells of 5 to 9 week old fetal tissue that resulted from elective abortions [13].

Box 1 | **Human Embryonic Stem Cells Colonies on Feeder layer**

Human embryonic stem cells are isolated by transferring the inner cell mass into a plastic laboratory culture dish that contains a nutrient broth known as cultured medium. The inner surface of the culture dish is typically coated with mouse embryonic skin cells that have been treated so they will not divide. This coating layer of cells is called a feeder layer. The reason for having the mouse cells in the bottom of the culture dish is to give the inner cell mass cells a sticky surface to which they can attach. Also, the feeder cells release nutrients into the culture medium. Recently, other ways of growing embryonic stem cells without the mouse feeder cells [4-6] have been established. This is a significant scientific advancement because of the risk that viruses or other macromolecules in the mouse cells may be transmitted to the human cells. Over the course of several days, the cells of the inner cell mass proliferate and begin to crowd the culture dish. When this occurs, they are removed gently and plated into several fresh culture dishes. Each cycle of sub culturing the cells is referred to as a passage. After six months or more, the original 30 cells of the inner cell mass yield millions of embryonic stem cells. Embryonic stem cells that have proliferated in cell culture for six or more months without differentiating, are pluripotent, and appear genetically normal, are referred to as an embryonic stem cell line (see Fig 1.3).

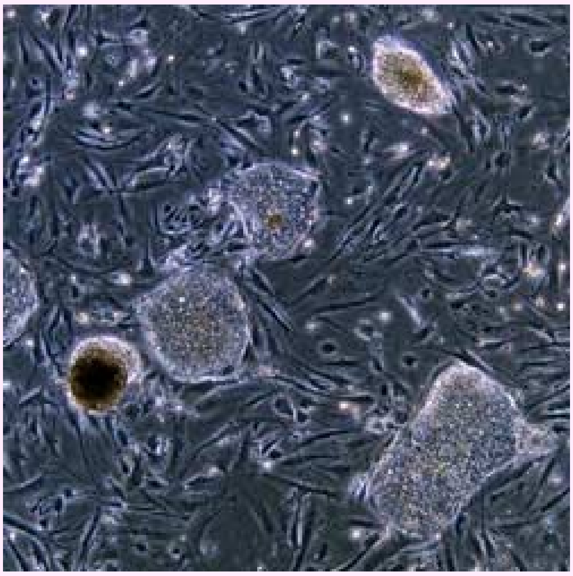


Figure 1.3 Human embryonic stem cell colonies on feeder layer.

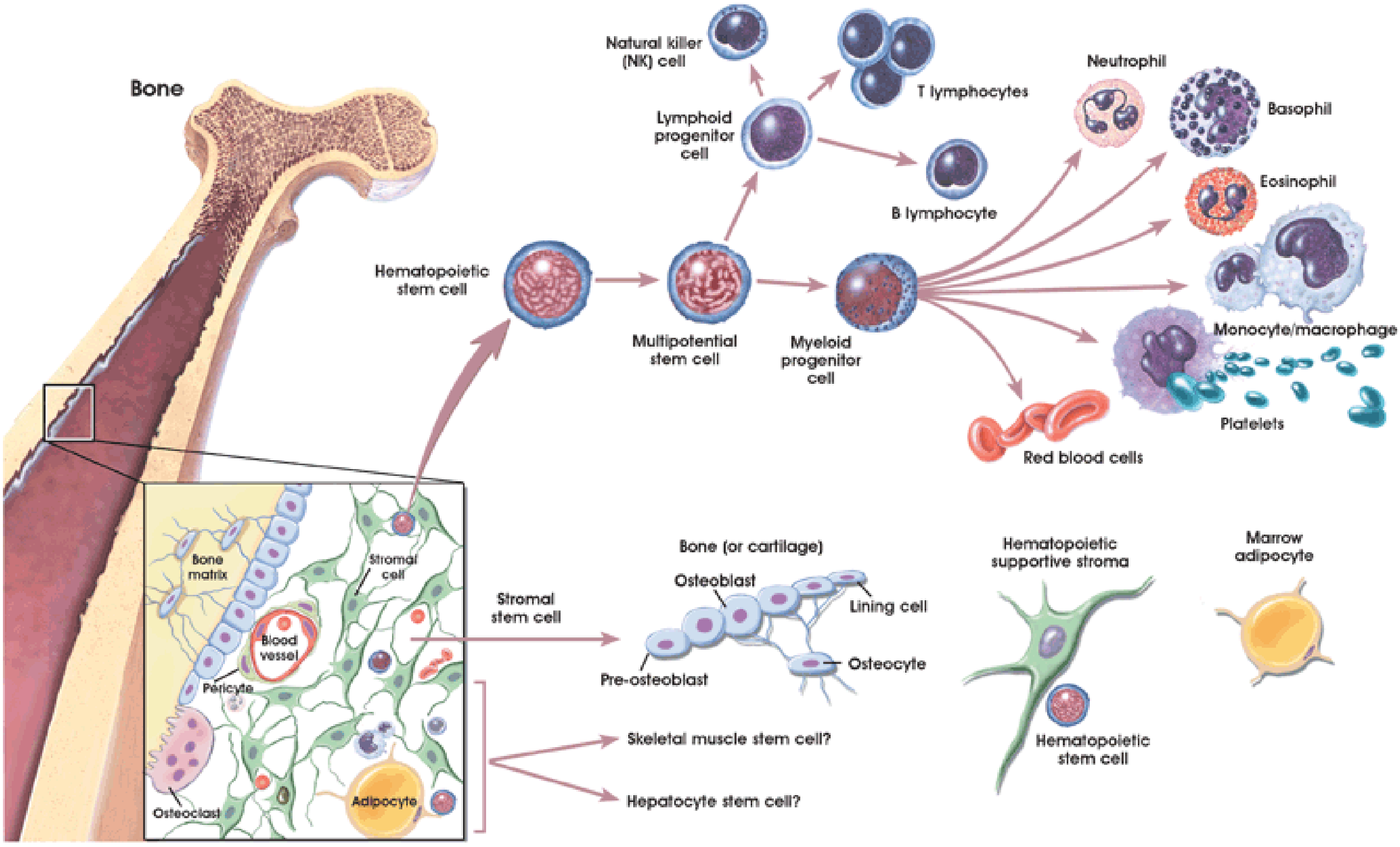
Laboratories that grow human embryonic stem cell lines use several kinds of tests. These tests include:

- Growing and sub culturing the stem cells for many months. This ensures that the cells are capable of long-term self-renewal.
- Using specific techniques to determine the presence of markers that are found only on undifferentiated cells like Oct-4. Oct-4 is a protein expressed by mouse and human ESC in vitro, and also by mouse inner cell mass in vivo. This protein and others (like Nanog) prevents differentiation. [14]
- Examining the chromosomes under a microscope. This is a method to assess whether the chromosomes are damaged or if the number of chromosomes has changed.
- Determining whether the cells can be sub cultured after freezing, thawing, and re-plating.
- Testing whether the human embryonic stem cells are pluripotent by 1) allowing the cells to differentiate spontaneously in cell culture; 2) manipulating the cells so they differentiate to form specific cell types; or 3) injecting the cells into an immunosuppressed mouse to test for the formation of a benign tumor called a teratoma [15-17]. Teratomas typically contain a mixture of many differentiated or partly differentiated cell types — indications that the embryonic stem cells are capable of differentiating into multiple cell types.

As long as the embryonic stem cells are grown in culture under appropriate conditions, they can remain undifferentiated (unspecialized). But if cells are allowed to clump together to form embryoid bodies, they begin to differentiate spontaneously [13]. To generate cultures of specific types of differentiated cells — heart muscle cells, blood cells, or nerve cells, for example — there is a need to control the differentiation of embryonic stem cells. This is done by changing the chemical composition of the culture medium, altering the surface of the culture dish, or modifying the cells by inserting specific genes. Through years of experimentation some basic protocols or "recipes" for the directed differentiation of embryonic stem cells into some specific cell types have been established.

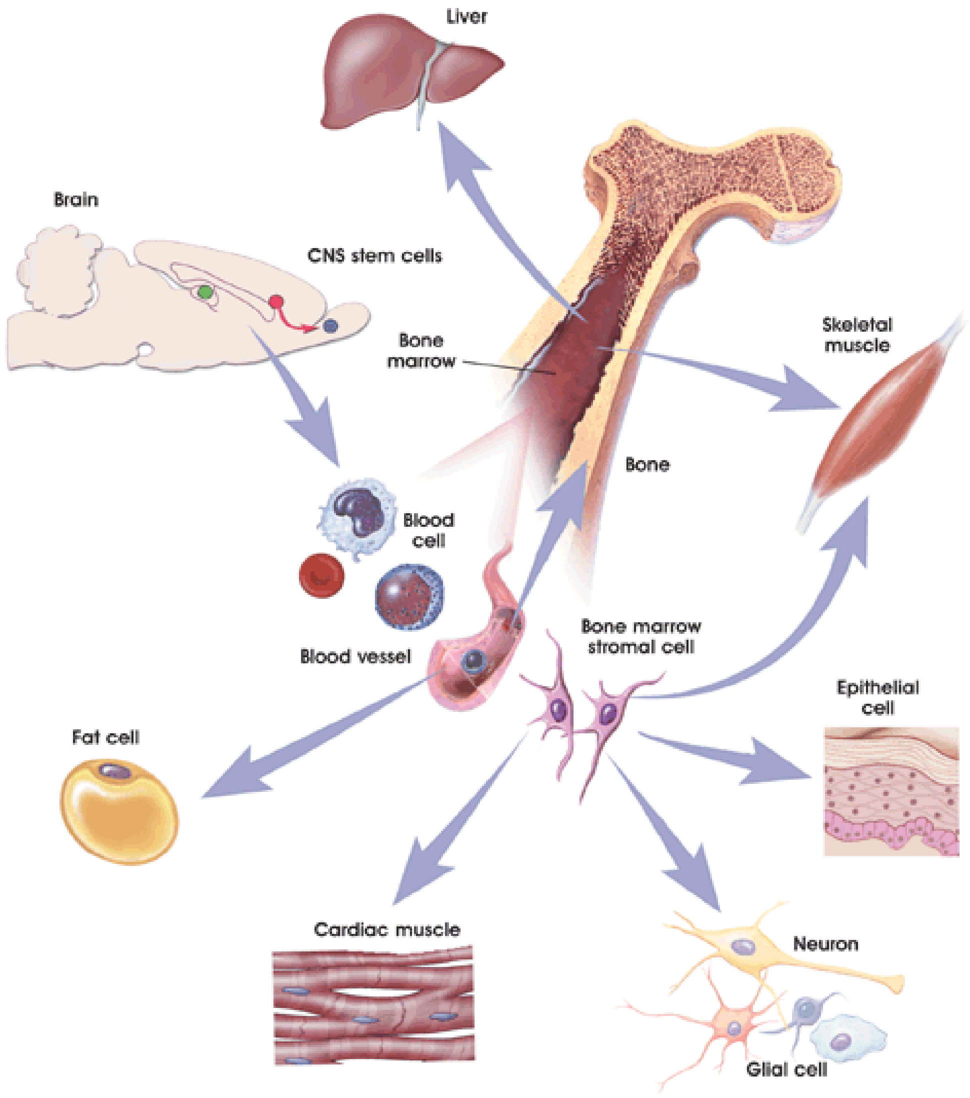
1.4 The Adult Stem Cells

An adult stem cell (ASC) is an undifferentiated (unspecialized) cell that occurs in a differentiated (specialized) tissue, can renew itself, and becomes specialized to yield all the specialized cell types of the tissue from which it originated. Adult stem cells are capable of making identical copies of themselves throughout the life time of the organism. Adult stem cells usually divide to generate progenitor or precursor cells, which then differentiate or develop into “mature” cell types that have characteristic shapes and specialized functions. Adult stem cells typically generate the cell types of the tissues in which they reside. A blood-forming adult stem cell in the bone marrow, for example, normally gives rise to the many types of blood cells such as red blood cells, white blood cells and platelets (see Fig. 1.4). Until recently it has been thought that a blood-forming cell in the bone marrow — which is called a hematopoietic stem cell (HSC)— could not give rise to the cells of a very different tissue, such as nerve cells in the brain. However, a number of experiments over the last several years have raised the possibility that stem cells from one tissue may be able to give rise to cell types of a completely different tissue, a phenomenon known as plasticity or transdifferentiation [18-21]. Examples of such plasticity include blood cells becoming neurons [18], bone marrow stem cells differentiate into another mesodermally derived tissue such as skeletal muscle [22, 23], heart muscle [24, 25] or liver [21, 26] (see Fig. 1.5). Therefore, exploring the possibility of using adult stem cells for cell-based therapies has become a very active area of investigation by researchers. Adult stem cells are rare. Their primary functions are to maintain the steady state functionality of a cell, called homeostasis, and with limitations, to replace cells that die because of injury and disease [7]. For example, only an estimated 1 in 10,000 to 15,000 cells in the bone marrow is a hematopoietic stem cell [27]. Furthermore, adult stem cells are dispersed in tissues throughout the mature animal and behave very differently, depending on their local environment. For example, HSC are constantly being generated in the bone marrow where they differentiate into mature types of blood cells. In contrast, stem cells in the small intestine are stationary, and are physically separated from the mature cells they generate.



© 2001 Terese Winslow, Lydia Kibiuk

Figure 1.4 Hematopoietic and Stromal Stem Cell Differentiations. Hematopoietic stem cells give rise to all the types of blood cells: red blood cells, B lymphocytes, T lymphocytes, natural killer cells, neutrophils, basophils, eosinophils, monocytes, macrophages, and platelets. Bone marrow stromal cells (mesenchymal stem cells) give rise to a variety of cell types: bone cells (osteocytes), cartilage cells (chondrocytes), fat cells (adipocytes), and other kinds of connective tissue cells such as those in tendons.



© 2001 Terese Winslow, Lydia Kibiuk, Caitlin Duckwall

Figure 1.5 Plasticity of adult stem cells The figure offers examples of adult stem cell plasticity that have been reported during the past few years. Hematopoietic stem cells may differentiate into three major types of brain cells (neurons, oligodendrocytes, and astrocytes); skeletal muscle cells; cardiac muscle cells; and liver cells. Bone marrow stromal cells may differentiate into cardiac muscle cells and skeletal muscle cells. Brain stem cells may differentiate into blood cells and skeletal muscle cells.

Many important questions about adult stem cells remain to be answered. They include:

- How many kinds of adult stem cells exist, and in which tissues do they exist?
- What are the sources of adult stem cells in the body? Are they "leftover" embryonic stem cells, or do they arise in some other way? Why do they remain in an undifferentiated state when all the cells around them have differentiated?
- Do adult stem cells normally exhibit plasticity, or do they only transdifferentiate when we manipulate them experimentally? What are the signals that regulate the proliferation and differentiation of stem cells that demonstrate plasticity?
- Is it possible to manipulate adult stem cells to enhance their proliferation so that sufficient tissue for transplants can be produced?
- Does a single type of stem cell exist — possibly in the bone marrow or circulating in the blood — that can generate the cells of any organ or tissue?
- What are the factors that stimulate stem cells to relocate to sites of injury or damage?

1.5 Comparison of Adult Stem Cells and Embryonic Stem Cells

Human embryonic and adult stem cells each have advantages and disadvantages regarding potential use for cell-based regenerative therapies. Of course, adult and embryonic stem cells differ in the number and type of differentiated cells types they can become. Embryonic stem cells can become all cell types of the body because they are pluripotent. Adult stem cells are generally limited to differentiating into different cell types of their tissue of origin. However, some evidence suggests that adult stem cell plasticity may exist, increasing the number of cell types a given adult stem cell can become.

Large numbers of embryonic stem cells can be relatively easily grown in culture, while adult stem cells are rare in mature tissues and methods for expanding their numbers in cell culture have not yet been worked out. This is an important distinction, as large numbers of cells are needed for stem cell replacement therapies.

A potential advantage of using stem cells from an adult is that the patient's own cells could be expanded in culture and then reintroduced into the patient. The use of the patient's own adult stem cells has the advantage that these cells are not rejected by the immune system. This represents a significant advantage, as immune rejection is a difficult problem that can only be circumvented with immunosuppressive drugs.

Embryonic stem cells from a donor introduced into a patient could cause transplant rejection. However, whether the recipient would reject donor embryonic stem cells has not been determined in human experiments.

For more information regarding stem cells:

Stem Cells: Scientific Progress and Future Research Directions.

Department of Health and Human Services, June 2001. http://www.nih.gov/news/stemcell/scireport.htm

1.6 Research Goals and Motivation

Despite the excitement surrounding stem cells' potential to perhaps cure disease or unlock the secret of development, a fundamental question remains: what, exactly, are stem cells? Although a few genes have been identified that seem to play a role in stem cells self-renewal, the key molecular switches remain a mystery. A year ago, two groups reported what they have hoped would be a significant step forward. As they described in papers published back to back in Science [28, 29], groups led by developmental geneticist Douglas Melton of Harvard University and Ihor Lemischka of Princeton University used gene chips to search for a common signal among different kinds of stem cells – a genetic profile that would in essence define the nature of "stemness" . Both Lemishcka and Melton found separate sets of genes that were over expressed in all stem cells. The problem was that the two sets of genes were completely different, sharing only six genes. Considering the identity of the experimental material and methods used in the two reports, it seems that "stemness" genes are elusive and cannot be readily identified by the approaches presented. These efforts have been made to identify a core program of "stemness" genes that account for both self renewal and pluripotency in *mouse* and are common to embryonic and adult stem cells.

Our work attempts to give an answer to this question looking into the genetic profile or transcriptomes of three stem cell tissues from *humans*: embryonic, hematopoietic and keratinocytic. A primary goal of our work was to identify how undifferentiated stem cells become differentiated. Turning genes on and off is central to this process. Some of the most serious medical conditions, such as cancer and birth defects, are due to abnormal cell division and differentiation. A better understanding of the genetic and molecular controls of these processes may yield information about how such diseases arise and suggest new strategies for therapy.

1.7 Research Plan

We have measured and analyzed stem cells' gene expression, starting with embryonic stem cells (ESC), which were derived from early embryo and are the source of all tissues during embryonal development. We further included adult stem cells from a variety of tissues which were recently suggested to also have a broad potential for differentiation as well as trans-differentiation, and were, therefore, candidates for tissue replacement therapy. Our first aim was to compare the genetic program of ESC and adult stem cells (ASC), in order to define their common expressed genes and to identify gene that are up-or down-regulated upon differentiation. We have used three sources of developmental and terminal differentiation stages of human cells: (i) embryonic stem cells (ESC), (ii) adult stem cells (ASC): hematopoietic (HSC) and keratinocytic (KSC), and (iii) their terminally differentiated counterparts (HDC and KDC).

RNA was extracted from each group of cells and processed for preparing targets for Affymetrix chips. A total of 17 hybridizations (samples) were performed in the experiment as follows:

ESC	HSC	HDC	KSC	KDC
3	4	4	3	3

We utilized a combination of *supervised statistical analysis* with *Super Paramagnetic Clustering (SPC)*, [30] a novel unsupervised clustering method for microarray data analysis. The analysis was aimed at defining common profiles of expression and to identify candidate genes involved in the different phases of the tissue differentiation. Furthermore, we intended to identify genes enriched in each individual stem cell population and then compare those sets of genes to one another (this work was done previously [28, 29] in mouse). This was done in order to search for new "stemness" genes in human stem cells, and in order to lead us to the understanding the genes that are responsible for pluripotency and to those that are turned off or on upon tissue differentiation.

Reference List

1. Rathjen, P.D., et al., *Properties and uses of embryonic stem cells: prospects for application to human biology and gene therapy.* Reprod Fertil Dev, 1998. **10**(1): p. 31-47.
2. Jones, J.M. and J.A. Thomson, *Human embryonic stem cell technology.* Semin Reprod Med, 2000. **18**(2): p. 219-23.
3. Thomson, J.A., et al., *Embryonic stem cell lines derived from human blastocysts.* Science, 1998. **282**(5391): p. 1145-7.
4. Amit, M., et al., *Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture.* Dev Biol, 2000. **227**(2): p. 271-8.
5. Itskovitz-Eldor, J., et al., *Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers.* Mol Med, 2000. **6**(2): p. 88-95.
6. Reubinoff, B.E., et al., *Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro.* Nat Biotechnol, 2000. **18**(4): p. 399-404.
7. Leblond, C.P., *Classification of Cell Populations on the Basis of Their Proliferative Behavior.* Natl Cancer Inst Monogr, 1964. **14**: p. 119-50.
8. Domen, J. and I.L. Weissman, *Self-renewal, differentiation or death: regulation and manipulation of hematopoietic stem cell fate.* Mol Med Today, 1999. **5**(5): p. 201-8.
9. Tosh, D. and J.M. Slack, *How cells change their phenotype.* Nat Rev Mol Cell Biol, 2002. **3**(3): p. 187-94.
10. Schuldiner, M., et al., *Effects of eight growth factors on the differentiation of cells derived from human embryonic stem cells.* Proc Natl Acad Sci U S A, 2000. **97**(21): p. 11307-12.
11. Brook, F.A. and R.L. Gardner, *The origin and efficient derivation of embryonic stem cells in the mouse.* Proc Natl Acad Sci U S A, 1997. **94**(11): p. 5709-12.
12. Evans, M.J. and M.H. Kaufman, *Establishment in culture of pluripotential cells from mouse embryos.* Nature, 1981. **292**(5819): p. 154-6.
13. Shamblott, M.J., et al., *Derivation of pluripotent stem cells from cultured human primordial germ cells.* Proc Natl Acad Sci U S A, 1998. **95**(23): p. 13726-31.
14. Chambers, I., et al., *Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells.* Cell, 2003. **113**(5): p. 643-55.
15. Kleinsmith, L.J. and G.B. Pierce, Jr., *Multipotentiality of Single Embryonal Carcinoma Cells.* Cancer Res, 1964. **24**: p. 1544-51.

16. Friedrich, T.D., U. Regenass, and L.C. Stevens, *Mouse genital ridges in organ culture: the effects of temperature on maturation and experimental induction of teratocarcinogenesis.* Differentiation, 1983. **24**(1): p. 60-4.
17. Andrews, P.W., *Human teratocarcinomas.* Biochim Biophys Acta, 1988. **948**(1): p. 17-36.
18. Brazelton, T.R., et al., *From marrow to brain: expression of neuronal phenotypes in adult mice.* Science, 2000. **290**(5497): p. 1775-9.
19. Krause, D.S., et al., *Multi-organ, multi-lineage engraftment by a single bone marrow-derived stem cell.* Cell, 2001. **105**(3): p. 369-77.
20. Anderson, D.J., F.H. Gage, and I.L. Weissman, *Can stem cells cross lineage boundaries?* Nat Med, 2001. **7**(4): p. 393-5.
21. Lagasse, E., et al., *Purified hematopoietic stem cells can differentiate into hepatocytes in vivo.* Nat Med, 2000. **6**(11): p. 1229-34.
22. Gussoni, E., et al., *Dystrophin expression in the mdx mouse restored by stem cell transplantation.* Nature, 1999. **401**(6751): p. 390-4.
23. Ferrari, G., et al., *Muscle regeneration by bone marrow-derived myogenic progenitors.* Science, 1998. **279**(5356): p. 1528-30.
24. Orlic, D., et al., *Bone marrow cells regenerate infarcted myocardium.* Nature, 2001. **410**(6829): p. 701-5.
25. Kocher, A.A., et al., *Neovascularization of ischemic myocardium by human bone-marrow-derived angioblasts prevents cardiomyocyte apoptosis, reduces remodeling and improves cardiac function.* Nat Med, 2001. **7**(4): p. 430-6.
26. Theise, N.D., et al., *Liver from bone marrow in humans.* Hepatology, 2000. **32**(1): p. 11-6.
27. Weissman, I.L., *Stem cells: units of development, units of regeneration, and units in evolution.* Cell, 2000. **100**(1): p. 157-68.
28. Ivanova, N.B., et al., *A stem cell molecular signature.* Science, 2002. **298**(5593): p. 601-4.
29. Ramalho-Santos, M., et al., *"Stemness": transcriptional profiling of embryonic and adult stem cells.* Science, 2002. **298**(5593): p. 597-600.
30. Blatt, M., S. Wiseman, and E. Domany, *Superparamagnetic clustering of data.* Physical Review Letters, 1996. **76**(18): p. 3251-3254.

Chapter 2

2 Materials and Methods

2.1 Samples used, extract preparation and labeling

Experiment Design:

This experiment is a comparative study of normal human cells at different stages of development and differentiation.

We have compared three developmental and terminal differentiation stages: (i) embryonic stem cells (ESC), (ii) adult stem cells (ASC): hematopoietic (HSC) and keratinocytic (KSC), and (iii) terminally differentiated counterparts (HDC and KDC).

A total of 17 hybridizations (samples) were performed in the experiment as follows:

ESC	HSC	HDC	KSC	KDC
3	4	4	3	3

No sample was used as a reference. Comparisons were made only between cell stages. At least three replicates, using either different biological samples or repeated hybridization, were performed for each cell stage.

Origin of the biological samples and their characteristics

All undifferentiated human embryonic stem cell (ESC) samples were obtained from the H9.2 clonal line (passages p29+40 - p29+58). This clone derives from the H9 human ES parent line, which was previously isolated from the inner cell mass of human blastocyst [1, 2] and approved by NIH (see figure 2.1). Both G-band and SKI assays showed that the H9.2 clonal line maintained a normal XX karyotype even after more than 8 months of continuous culture [1, 2].

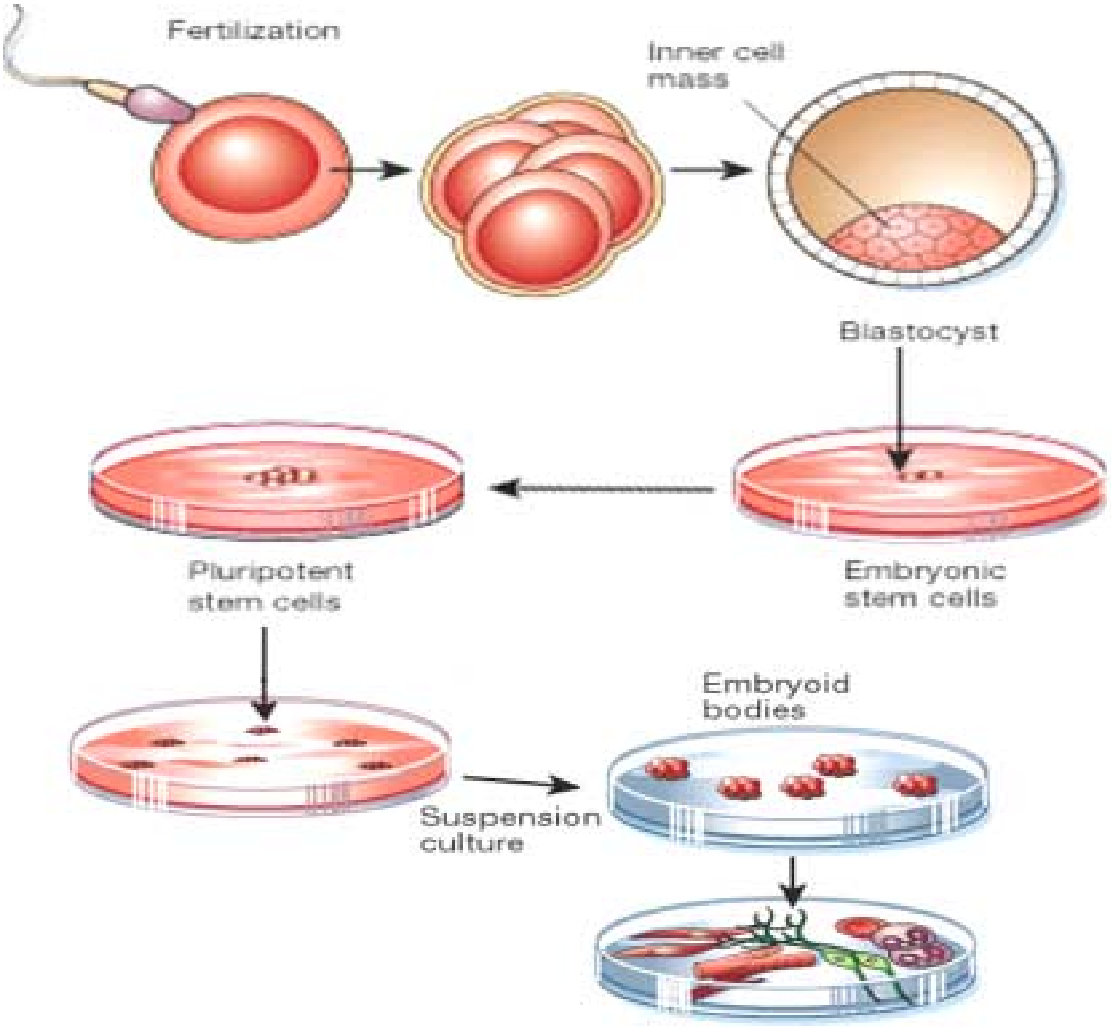


Figure 2.1 Origin of human Embryonic Stem Cell (ESC). ESC are derived from the inner cell mass of the pre-implantation embryo [1, 2]. Differentiation can be induced by growing of stem cell colonies in suspension culture to form Embryoid Bodies cells (EBC), which upon dissociation can be plated to yield differentiating cells.

Hematopoietic cells were obtained from (i) 2 pools (5 units and 15 units, 75 ml/unit) of cord blood collected after placental separation according to routine procedure approved by Institutional Review Board (IRB), and from (ii) peripheral blood collected by pheresis from adult normal donors primed with four daily injections of G-CSF (10 µg/kg/day), using the Cobspectra stem cell collector.

Keratinocyte cells were obtained from (i) 12 pooled neonatal foreskins of 8 days old donors after ritual circumcision and informed consent of the parents. All epidermal cells were isolated from the epidermal tissue as previously described [3]. Alternatively (ii), cells were obtained from primary cultures of normal human epidermal keratinocytes, previously isolated as described above and further sub cultured as described below.

Manipulation of biological samples and protocols used for growth conditions and separation techniques

Non-differentiating ESC lines H9.2 were grown on an inactivated mouse embryonic feeder layer (37°C, 5% CO2) (MEF) [1, 2]. Cells were grown in a culture medium consisting of 80% KO-DMEM, supplemented with 20% SR, 1 mM L-glutamine, 0.1 mM β–mercaptoethanol, 1% non-essential amino acid stock, and 4 ng/ml bFGF (Gibco Invitrogen, San Diego, CA). They were passaged every four to six days using 1 mg/ml type IV collagenase treatment (Gibco Invitrogen, San Diego, CA). Immortality and pluripotency were verified by in vitro expression of specific primate embryonic markers such as telomerase activity, OCT4, SSEA4, TRA1-60 and TRA1-81, and by in vivo teratoma formation after injection into the hind limb muscle of SCID mice as previously described [1, 2]. ESC were separated from the feeder layer by type IV collagenase treatment as described above followed by microscopical inspection for the absence of contamination by feeder cells (3 samples). Once separated and removed from the feeder layer, about 106 cells were injected into the hind limb muscle of 4-week-old male SCID beige mice (Harlan, Israel). Teratomas could be detected after 4 weeks and were removed for histological and immunohistochemical examination at least 10 weeks after the injection.

Hematopoietic cord blood cells (HSC, 2 samples) were subjected to ficoll gradient and the cells were enriched using anti CD133 magnetic beads separation system (Miltenyi) (see figure 2.2). Hematopoietic peripheral blood cells (HSC, 2 samples) were enriched using anti CD133 magnetic beads system. The yield of CD133 positive cells was 0.14% for cord blood and 0.7% for peripheral blood and the isolated cell populations were 80-85% positive for CD133 as assayed by FACS. The non-selected cells from cord or peripheral blood served as differentiated cells and were termed HDC (4 samples).

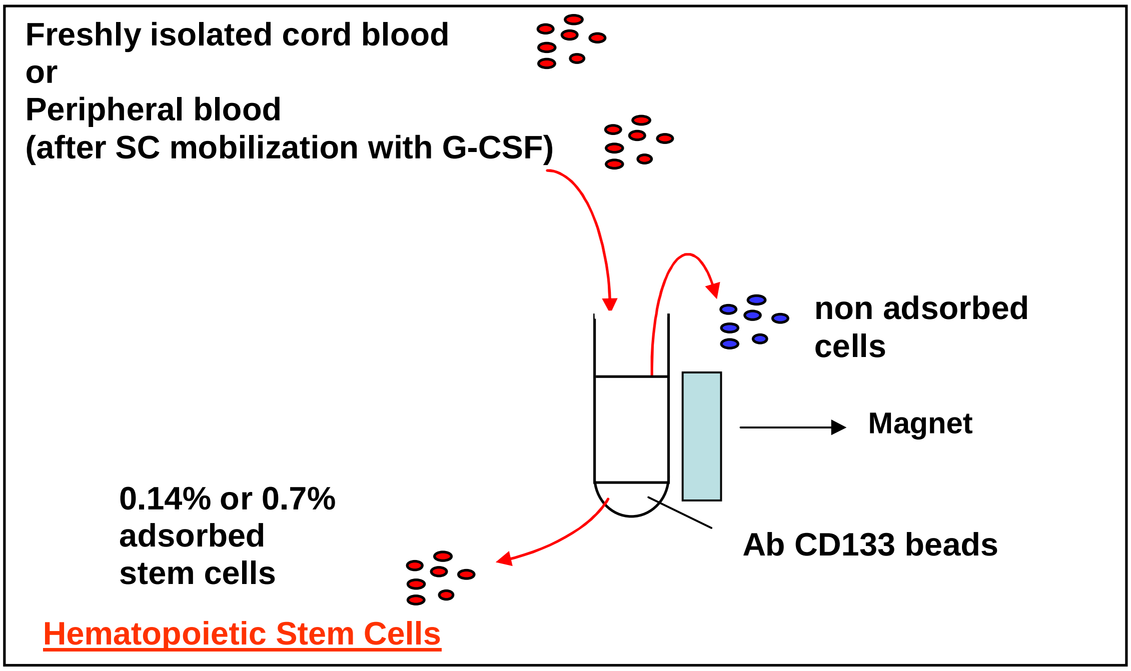


Figure 2.2 Isolation of human Hematopoietic Stem Cell (HSC). Hematopoietic cells were obtained from (i) 2 pools of cord blood collected after placental separation, and from (ii) peripheral blood collected by pheresis from adult normal donors primed with four daily injections of G-CSF. Using anti CD133 magnetic beads separation system (Miltenyi), the yield of CD133 positive cells was 0.14% for cord blood and 0.7% for peripheral blood and the isolated cell populations were 80-85% positive for CD133 as assayed by FACS. The non-selected cells from cord or peripheral blood served as differentiated cells and were termed Hematopoietic Differentiated Cells (HDC).

To allow proliferation without favoring differentiation, isolated keratinocytes were co-cultured in Keratinocyte Growth Medium (KGM) (37°C, 5% CO2) in the presence of mitomycin C treated-feeder layer of mouse fibroblasts as previously described [3]. KGM consists of a mixture

(3:1) of DMEM and Ham F12 (Gibco Invitrogen, San Diego, CA), enriched with adenine (1.8 x 10-4 M), insulin (5 µg/ml), the HCE cocktail (Sigma, St Louis, MO): hydrocortisone (0.4 µg/ml), cholera toxin (0.1 nM), EGF (20 ng/ml), and supplemented with 10% FBS (Gibco Invitrogen, San Diego, CA). The culture media were changed every 2 to 3 days until cells reached 80% confluence, after what cells were further sub cultured for a maximum of 2 passages. The J2-3T3 feeder cell line is a clone derived from NIH 3T3 cells and selected for their efficiency at supporting keratinocyte growth. J2-3T3 cells were maintained in DMEM, supplemented with 10% Donor Calf Serum (Gibco Invitrogen, San Diego, CA). Keratinocyte stem cells were enriched by differential adsorption of low-passaged (≤ 2) cultured human keratinocytes or freshly isolated neonatal foreskin keratinocytes on type IV collagen coated plates as adapted from Jones et al. [4]. Feeder layer cells were removed from cultured keratinocytes by rapid (5') treatment with trypsin, followed by washes of PBS to remove all the feeder cells. Adherent cultured keratinocytes were checked by microscopical inspection for the absence of feeder contamination and further harvested after prolonged (>20') trypsin treatment. Rapidly adherent cells (progenitor "stem cells") were harvested after ≤ 1h adsorption and termed KSC (3 samples). Unadsorbed cells (Transit Amplifying Cells, TAC) were collected and plated again overnight on other type IV collagen coated plates. The remaining unadsorbed cells (terminally differentiated cells) were collected and termed KDC (3 samples) (see figure 2.3). The yield of KSC was less than 0.4% of the isolated epidermal or cultured cells. The isolated KSC and their committed (TAC) and differentiated counterparts (KDC) were characterized by clonogenicity assay and expression of various specific markers (Figure 2.4) [3, 5].

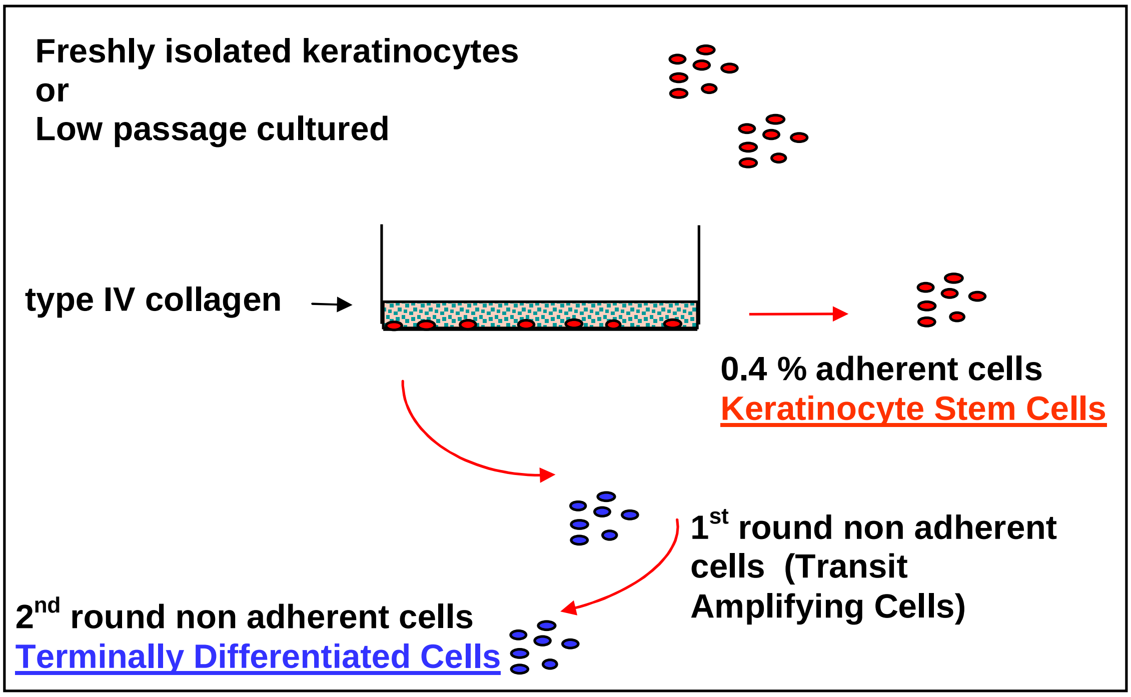

Freshly isolated keratinocytes
or
Low passage cultured
type IV collagen
0.4 % adherent cells
Keratinocyte Stem Cells
1st round non adherent cells (Transit Amplifying Cells)
2nd round non adherent cells
Terminally Differentiated Cells

Figure 2.3 Isolation of human Keratinocytes Stem Cell (KSC). Keratinocyte cells were obtained from (i) 12 pooled neonatal foreskins of 8 days old donors after ritual circumcision and informed consent of the parents. Alternatively (ii), cells were obtained from primary cultures of normal human epidermal keratinocytes, previously isolated as described above and further sub cultured as described below. All epidermal cells were isolated from the epidermal tissue as previously described [3]. Keratinocyte stem cells were enriched by differential adsorption of cultured human keratinocytes or freshly isolated neonatal foreskin keratinocytes on type IV collagen coated plates as adapted from Jones *et al.* Rapidly adherent cells (progenitor "stem cells") were harvested after ≤ 1h adsorption and termed KSC. Unadsorbed cells (Transit Amplifying Cells, TAC) were collected and plated again overnight on other type IV collagen coated plates. The remaining unadsorbed cells (terminally differentiated cells) were collected and termed KDC. The yield of KSC was less than 0.4% of the isolated epidermal or cultured cells.

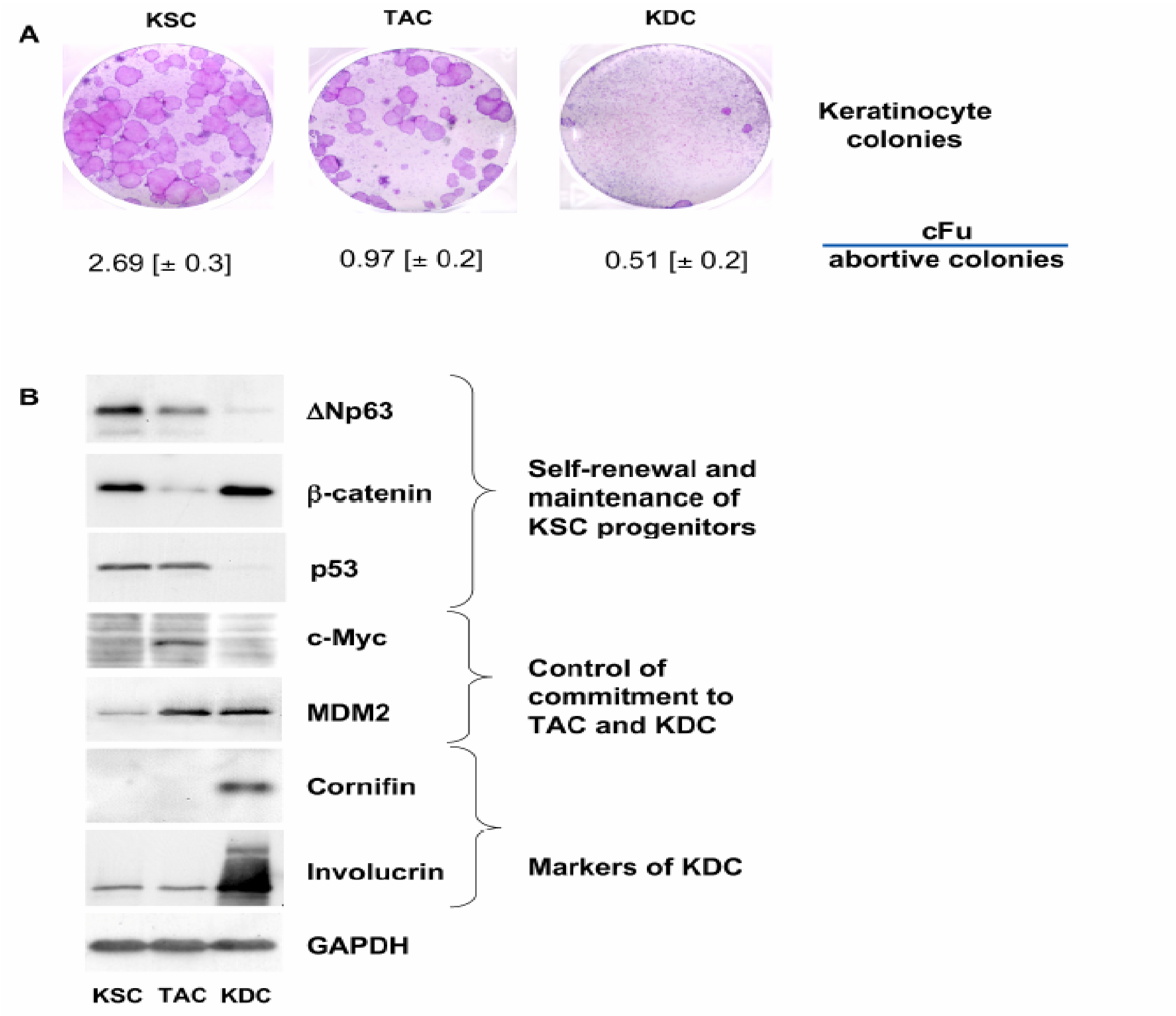


Figure 2.4 Clonogenicity assay and expression profile of epidermal specific markers of keratinocyte fractions. A, Clonogenicity assay: after selection of isolated keratinocytes from human epidermis on type IV collagen (see figure 3), 2000 cells of each fraction were plated per well (in triplicate), and after two weeks in culture, keratinocyte colonies were scored. Numbers represent averaged cFu / abortive colonies. Standard errors are in brackets [3-5]. **B**, Western blot analysis of specific markers in keratinocyte fractions isolated as in a. KSC, Keratinocyte Stem Cells; TAC, Transit Amplifying Cells; and KDC, terminally differentiated keratinocytes.

Protocols for preparing the hybridization extracts

Total RNA was extracted from each sample using total RNA isolation reagent TRIzol® (Gibco Invitrogen™, San Diego, CA) with minor modifications from the manufacturer's recommendations (http://www.invitrogen.com/content/sfs/manuals/15596026.pdf).

The amount of starting RNA was determined by UV absorption using a RNA/DNA calculator (GeneQuant™, Amersham Biosciences, Piscataway, NJ), and the quality of RNA was assessed on agarose gel. Total RNA from each sample was used to prepare biotinylated target cRNA, according to Affymetrix™ manufacturer's recommendations

10 µg of total RNA was used to generate first-strand cDNA by using a T7-linked oligo(dT) primer. After second-strand synthesis, in vitro transcription was performed with biotinylated UTP and CTP (BioArray™ HighYield™ RNA transcript labeling kit, Enzo Life Sciences, Farmingdale, NY), resulting in approximately 100-fold amplification of cRNA.

External (spikes) and internal controls

Target cDNA generated from each sample were processed as per manufacturer's recommendation using an Affymetrix GeneChip® Instrument System http://www.affymetrix.com/support/technical/manual/expression_manual.affx. Spike controls were added to 10 µg fragmented cRNA before each sample hybridization.

Housekeeping Controls:	**Spike Controls:**
HUMISGF3A / M97935	BIOB
HUMRGE / M10098	BIOC
HUMGAPDH / M33197	BIODN
HSAC07 / X00351	CREX
M27830	

3'/5' ratios for GAPDH and beta-actin were confirmed to be within acceptable limits (0.85-1.63), and BioB spike controls were found to be present on all chips, with BioC, BioD and CreX also present in increasing intensity. When scaled to a target intensity of 150 (using Affymetrix MAS 5.0 array analysis software, see below), scaling factors for all arrays were within acceptable limits (0.86-1.26 fold), as were background, Q values and mean intensities.

Hybridization procedures and parameters:

An Affymetrix test chip (TEST 3), containing approximately 350 genes, was run prior to each sample on the original HG-U133A to check for target cRNA integrity and labeling and good quality of aforementioned controls. Hybridizations were performed at 45°C for 16h. Arrays were then washed and stained with streptavidin-phycoerythrin, further amplified with biotinylated - anti streptavidin and stained again with streptavidin-phycoerythrin.

Measurement data and specifications:

Arrays were scanned by Affymetrix™ GeneChip® scanner. Raw data images (.DAT file) were generated and analyzed by MAS 5.0 Affymetrix™ array analysis software. After scanning, array images were assessed by eye to confirm scanner alignment and the absence of significant bubbles or scratches. The files, which contain the average intensity of each probe cell (.CEL file), were automatically generated from the DAT files by MAS 5.0 software.

Array Design

Antisense biotinylated target cRNA were hybridized to an in situ synthesized oligonucleotide microarray (see figure 2.5) HG-U133A GeneChip® Affymetrix™

(http://www.affymetrix.com/products/arrays/specific/hgu133.affx).

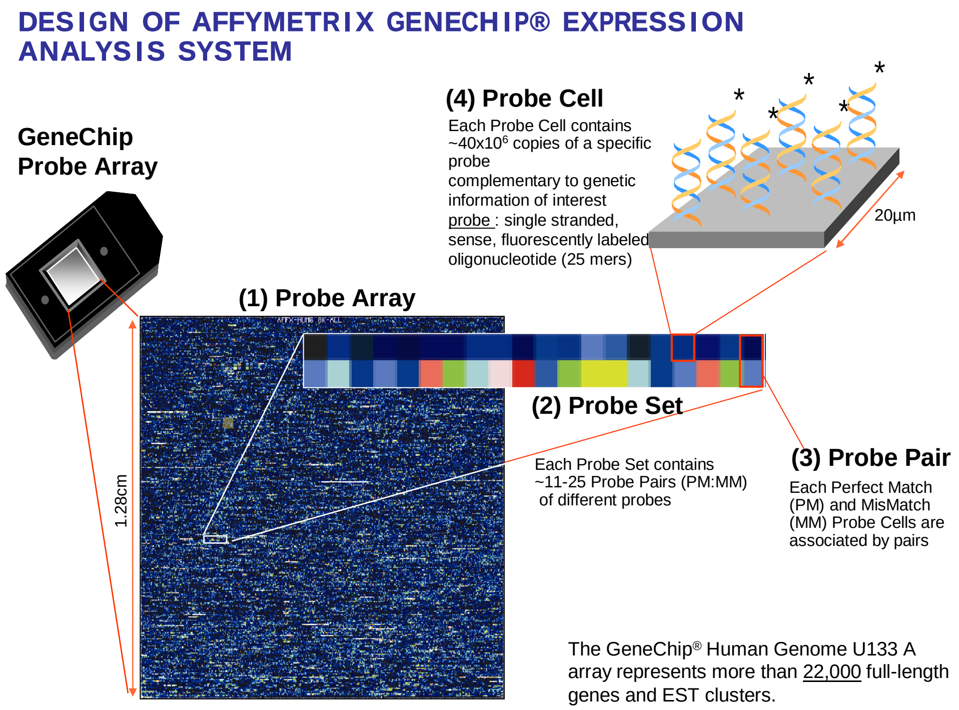

DESIGN OF AFFYMETRIX GENECHIP® EXPRESSION ANALYSIS SYSTEM
GeneChip Probe Array
(4) Probe Cell
Each Probe Cell contains ~40x10^6 copies of a specific probe complementary to genetic information of interest
probe : single stranded, sense, fluorescently labeled oligonucleotide (25 mers)
20µm
(1) Probe Array
1.28cm
(2) Probe Set
Each Probe Set contains ~11-25 Probe Pairs (PM:MM) of different probes
(3) Probe Pair
Each Perfect Match (PM) and MisMatch (MM) Probe Cells are associated by pairs
The GeneChip® Human Genome U133 A array represents more than 22,000 full-length genes and EST clusters.

Figure 2.5 Design of Affymetrix GeneChip® Expression Analysis System. **(1)** Probe array is the chip containing around 22,000 probe sets (genes or EST). **(2)** Probe set is a set of probes designed to detect one transcript. A probe set usually consists of 16-20 probe pairs. **(3)** Probe pair is two probe cells, a PM and its corresponding MM. **(4)** Probe cell is a single square–shaped feature on an array containing one type of probe. Each probe cell contains millions of probes molecules. Probe is a single 25 base long stranded DNA oligonucleotide complementary to a specific sequenece.

Calculation Gene Expression

For probe-level data analysis, we tried several methods of probe set summarization on all the original CEL files:

MBEI (http://www.biostat.harvard.edu/complab/dchip/)

RMA (http://www.bioconductor.org)

MAS-5.0 (http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf).

MAS 5.0 and MBEI gave similar results. However, RMA washed out the biologically relevant differences between stem and non stem samples (see figure 3.1 in Results, chapter 3). According to our checking the discrepancy lies in the *quantile normalization* process which is a part of the RMA algorithm.

Quantile Normalization

The goal of quantile normalization is to make the distribution of probe intensities (just the PM) the same for all the chips $i = 1....N$. This approach is based upon the assumption that the distribution of intensities for each chip should be the same.

1. Given N chips of length P (usually $20\times\#$ probe sets on the chip) that form a matrix X of dimension $P \times N$.
2. Set $d = (\frac{1}{\sqrt{N}} ...... \frac{1}{\sqrt{N}})$
3. Sort each column of X to give X_{sort}.
4. Project each row of X_{sort} onto d to get X'_{sort} - The projection is equivalent to taking the average of a particular row and substituting this value for each of the individual elements in that row. If $q_i = (q_{i1} ...... q_{iN})$ is a row in X_{sort} then the corresponding row in X'_{sort} is given by $q'_i = proj_d q_i$.
5. Get X_{norm} by rearranging each column of X'_{sort} to have the same ordering as the original X.
6. The signal of each probe set is calculated using the X_{norm} values. Obviously, the distributions of the elements of X_{norm} in every column are identical.

We have decided to use MAS 5.0 expression values because it gave similar results to MBEI and to RMA without quantile normalization.

Calculation Gene Expression by MAS 5.0 Affymetrix™

The main software from Affymetrix is MicroArraySuite-MAS version 5 (MAS 5.0). The output of this software consists of the following files:

- EXP file: contains the meta-data about the experiment including name of researcher, name of experiment, sample type, name and type of GeneChip, target synthesis- hybridization and washing protocols.
- DAT file: An image file, scanned GeneChip image at the pixel level (~10^7 pixels, ~50 MB).
- CEL file: Cell intensity file, probe level PM and MM values.
- CDF file: Chip Description File. Describes which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs).
- CHP file: Analyzed cell intensities (e.g. after MAS 5.0).
- RPT file: report file.

The CEL file has been computed using the DAT file in the following way: Each probe cell in the CEL file contains 10x10 pixels. In order to calculate the probe cell signal (PM or MM) the algorithm removes the outer 36 pixels and computes the 75th percentile (taking from the probe cells distribution only the 75th percent or below) of the 8x8 pixel values of each probe. Furthermore, from each probe cell signal value one subtracts by the background noise, which is the average of the lowest 2% probe cells in its sector. (Usually, the probe array is divided into 16 sectors). This way the CEL file is generated.

PMij, MMij= Intensity for Perfect Match and MisMatch probe pair i in probe set j.

i = 1,..., I--usually 16 or 20 probe pairs;

j = 1,..., J--between 8,000 and 20,000 probe sets.

The Detection Algorithm: Detection p-value is used to assign Present, Marginal and Absent calls to genes.

Probe pairs were scored for their ability to detect targets through the Discrimination Score R which reflects the ability of a probe pair to hybridize to its target transcript. R is the ratio of target specific intensity (PM-MM) and the total hybridization intensity (PM+MM).

$$R = \frac{PM - MM}{PM + MM}$$

Detection p-value were calculated using Wilcoxson's Signed Rank Test for the R values that lie within the default discrimination threshold (τ = 0.015).

To make a Presence, Marginal or Absence call, the detection p-value were compared within pre-set boundaries ($\alpha 1$ = 0.04 and $\alpha 2$ = 0.06).

Wilcoxson's Signed Rank Test on the difference d=R- τ works as follows:

1. Null hypothesis (H0): d <=0
2. Rank all the probes within a probe set, by the absolute values of d and then set the sign of each rank to the sign of the corresponding d.
3. T = Sum all the positive rank values
4. Calculate probability of exceeding the Rank Sum Score of T

   $$P = \frac{Combination > T}{TotalCombination}$$

5. Normally we would reject the H0 (which means most of the d for the probe set are positive) if p<0.05. Affymetrix sets two thresholds, 0.04 and 0.06: The call is Present if p<0.04.
6. If the p-value is bigger than 0.06 THEN d<=0. In this case the call is absent
7. A very few probe sets will have p-value between 0.04 and 0.06. These are marginal calls

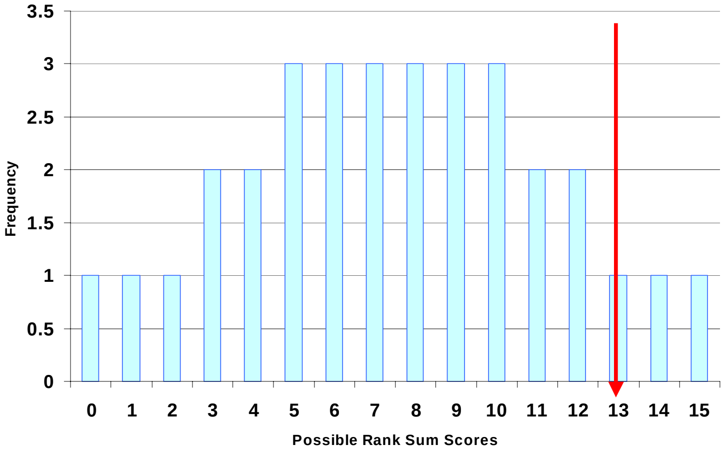
For example, consider the following hypothetical probe set:

PM	MM	R	tau	d	Absolute d	Rank	Signed Rank
5000	1000	0.6667	0.015	0.6517	0.6517	5	5
4000	1000	0.6000	0.015	0.5850	0.5850	4	4
3000	1000	0.5000	0.015	0.4850	0.4850	3	3
2000	1000	0.3333	0.015	0.3183	0.3183	1	1
500	1000	-0.3333	0.015	-0.3483	0.3483	2	-2

T = Sum of the positive rank = 13

What is the probability of exceeding the Rank Sum Score of 13?

The distribution of rank sums:



#combinations with Rank Sum of 13 =1

#combinations with Rank Sum > 13 = 2

Total combinations = 2^5 = 32

$$P = \frac{1*0.5+2*1}{2^5} = 0.0781$$

The Signal Algorithm: computing the .CHP file

The Signal is a value that reflects the relative abundance of a transcript. Each probe pair contributes to the final signal value. If MM < PM then the MM is considered informative and used as an estimate of background (stray) signal. If MM are generally informative except a few, then those are replaced by an adjusted MM value. If the MM values are generally uninformative (MM > PM) they are replaced by values that are slightly smaller than PM (such probe sets more often than not, receive Absent calls). To calculate a Specific Background (SB_i) ratio representative for the probe set, we use the One-step Tukey's Biweight algorithm (see below). We find a typical log ratio of PM to MM that is simply an estimate of the difference of log intensities for a selected probe set. The Biweight Specific Background (SB_i) for probe pair j in probe set i is:

$$SB_i = T_{bi}\left(\log_2(PM_{i,j}) - \log_2(MM_{i,j}) : j = 1,\ldots,n_i\right)$$

If SB_i is large, then the values from the probe set are generally reliable, and we can use SB_i to construct the Ideal Mismatch (IM) for a probe pair if needed. If SB_i is small (SB_i <*contrastτ*), we smoothly degrade to use more of the PM value as the Ideal Mismatch. The three cases of determining IM for probe pair in probe set i are described in the following formula:

$$IM_{i,j} = \begin{cases} MM_{i,j}, & MM_{i,j} < PM_{i,j} \\ \dfrac{PM_{i,j}}{2^{(SB_i)}}, & MM_{i,j} \geq PM_{i,j} \text{ and } SB_i > contrast\tau \\ \dfrac{PM_{i,j}}{2^{\left(\frac{contrast\tau}{1+\left(\frac{contrast\tau - SB_i}{scale\tau}\right)}\right)}}, & MM_{i,j} \geq PM_{i,j} \text{ and } SB_i \leq contrast\tau \end{cases}$$

default $contrast\tau = 0.03$

default $scale\tau = 10$

The first case where the mismatch value provides a probe-specific estimate of stray signal is the best situation. In the second case, the estimate is not probe-specific, but at least provides information specific to the probe set. The third case involves the least informative estimate, based only weakly on probe-set specific data.

The signal probe value (PV) is calculated by a weighted mean of probe fluorescence (corrected for non specific signal by subtracting the Ideal Mismatch (IM) probe value) using again the One-step Tukey's Biweight Estimate.

$$V_{i,j} = \max((PM_{i,j} - IM_{i,j}), \delta)$$

$$PV_{i,j} = \log_2(V_{i,j}),\ j = 1,\ldots,n_i$$

where n is the number of probe pairs in the probe set and default δ=2^{-20}

$$SignalLogValue = T_{bi}(PV_{i,1,\ldots,PV_{i,n_i}})$$

One-Step Tukey's Biweight Algorithm is used to calculate a robust average - A median is computed to define the center of the data. The distance of each data point from the median determines the extent to which it contributes to the Signal (this decreases the influence of outliers with extremely low or high values). This Signal value, a relative measure of the expression level, was computed for each assayed gene.

Scaling Factor: If the algorithm settings indicate scaling all probes sets to a target, we calculate a scaling factor (sf)

$$Sf = \frac{Sc}{TrimMean(SignalValues, 0.02, 0.98)}$$

Where Sc is the target signal (in our case Sc = 150). The TrimMean function here takes the average value of all observations after removing the values in the lowest 2% of observations and removing those values in the upper 2% of observations.

The reported value of probe set i is:

$$SacledSignal = sf * signal$$

For more information see http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

2.2 Analysis Methods

Preprocessing and filtering

First, 15427 probe sets with at least one “present” call were selected. Expression levels < 30 were thresholded to 30 and log2 was taken to generate the final gene expression matrix (17 x 15427). We analyzed two groups of samples:

(A) hematopoieitic (H) pathway, ESC -> HSC -> HDC (3+4+4 samples) and

(B) keratinocytic (K) pathway ESC -> KSC -> KDC (3+3+3 samples).

For each group the genes were filtered using *ANOVA* [6]. False discovery rate (FDR) [7] was controlled at 0.05. This left 8290 PS (6293 genes) that vary significantly over the three kinds of cell states in group (A) and 5432 (4301 genes) for group (B). The smaller number of genes for the K-pathway reflects the smaller number of samples.

Normalization Prior to Clustering

Before clustering the rows of the data matrix (genes) are centered (mean=0) and normalized to standard deviation of 1:

$$E'_{gs} = \frac{E_{gs} - \overline{E}_g}{\sqrt{\sum_s (E_{gs} - \overline{E}_g)^2}}$$

2.3 One Way ANOVA

A One-Way Analysis of Variance [6] is a way to test the equality of three or more means at one time by using variances.

Assumptions:

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.

Hypotheses:

The null hypothesis will be that all population means are equal; the alternative hypothesis is that at least one mean is different.

The whole idea behind the analysis of variance is to compare the ratio of between group variance to within group variance. If the variance caused by the difference between the groups is much larger than the variance within each group, we conclude that the means aren't the same.

Grand Mean ($\overline{X}_{GM}$): the grand mean of a set of samples is the total of all the data values divided by the total sample size (N).

$$\overline{X}_{GM} = \frac{\sum X_{ij}}{N}$$

Total Variation (SS_T): the total variation is comprised the sum of the squares of the differences of each value with the grand mean.

$$SS_T = \sum\sum (X_{ij} - \overline{X}_{GM})^2$$

Between Group Variation (SS_B): the variation due to the interaction between the samples is the Sum of Squares between groups.

$$SS_B = \sum_j n_j (\overline{X}_j - \overline{X}_{GM})^2$$

where n_j is the number of samples in group j and $\overline{X}_j$ is the mean of the set of samples in group j.

Within Group Variation (SS_W): the variation due to differences within individual samples is the Sum of Squares Within groups.

$$SS_w = \sum\sum(X_{ij} - \bar{X}_j)^2$$

Let's denote k is the number of groups, we can summarize it by the following table:

	SS	**df**	**MS**	**F**
Between	SS_B	k-1	SS_B ----------- k-1	MS_B -------------- MS_W
Within	SS_W	N-k	SS_W ----------- N-k	
Total	$SS_W + SS_B$	N-1		

F test statistic: the F test statistic is found by dividing the between group variance by the within group variance. The decision will be to reject the null hypothesis if the test statistic from the table is greater than the F critical value with k-1 numerator and N-k denominator degrees of freedom.

If the decision is to reject the null, then at least one of the means is different. However, the ANOVA does not tell you where the difference lies.

2.4 Controlling the False Discovery Rate (FDR)

The Multiplicity Problem

DNA microarrays have been used for the purpose of monitoring expression levels of thousands of genes simultaneously and identifying those genes that are differentially expressed. The probability that a false identification (type I error) is committed can increase sharply when the number of tested genes gets large. Correlation between the test statistics attributed to gene co-regulation and dependency in the measurement errors of the gene expression levels further complicates the problem.

The False Discovery Rate (FDR) [7]

The multiplicity problem was originally addressed by methods to control the family-wise type I error rate (FEW) which is the probability of committing at least one error in the family of hypotheses. A simple example of FEW is the Bonferroni method. Using this method, we reject the null hypothesis only in cases where $p < \frac{\alpha}{N}$, N being the number of tests preformed. This insures that the expectancy of false positives is α, and thus the probability to get even one false positive is less than α.

In DNA microarray experiments, the number of tests preformed is in the order of thousands. Therefore, a method such as Bonferroni will require very small p-values and will result in a significant loss of power. As an alternative, one can supply a measure for the expected proportion of falsely discovered genes among the list of genes that are identified; the expected proportion is the FDR.

The Procedure

Let N be the number of null hypotheses tested. For each hypothesis H_g, a test statistic is calculated with a corresponding p-value, p_g.

The *N* genes are ordered according to their p_g values. An upper bound, *q*, for the fraction of false positives is set; and the minimal index, j, for which $p_i > i \times {}^{q}\!/_{N}$ is found for all $i > j$. The null hypothesis is rejected for all genes with index $i \le j$. At the end of this procedure we are left with a list of genes for which the expected fraction of false positives is q.

2.5 Tissue Specific Analysis (Z-score)

The GNF dataset (http://expression.gnf.org/cgi-bin/index.cgi) of Su *et al.* [8], supplemented by four measurements of expression in keratinocytes [5], was used to determine tissue specific expression of various genes. We performed MAS 5.0 analysis on all the original CEL files of these 4 samples and those of the GNF dataset. For each of the Nc genes of a cluster c we found the matching probe set in the U95 chip using Unigene (build #158) and GenBank® accession numbers; we refer to this dataset as GNF*. We used expression values characteristic of 21 tissues, obtained by averaging the results of several repeats and sub-types. We performed, for each gene g, 21 statistical Z-score tests, to determine whether g is expressed at a higher level in tissue i than in the other 20 tissues. Lets denote the expression level of a gene g in tissue i by Y_{gi}, and in the other 20 tissues as X_{gi}. The Z-score is:

$$Z_{gi} = \frac{Y_{gi} - mean(X_{gi})}{std(X_{gi})}$$

For each tissue i we calculated P-values for N_c genes,

$$P_{gi} = 1 - normcdf(Z_{gi})$$

and prepared, using FDR of 0.05, a list of genes whose expression level is specific to the tissue M_{ci}.

2.6 Chi-Square Test for Independence

Chi-square [9] is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. The chi-square test is testing the null hypothesis, which states that there is no significant difference between the expected and observed result. Chi-square determines the independence of the rows and columns of the table according to the following steps:

1. Create a table of cell frequencies. Compute row and column totals.

	1	2	Total
1	O_{11}	O_{12}	$O_{11} + O_{12} = R_1$
2	O_{21}	O_{22}	$O_{21} + O_{22} = R_2$
Total	$O_{11} + O_{21} = C_1$	$O_{12} + O_{22} = C_2$	T

2. Compute expected cell frequencies using the formula:

$$E_{ij} = \frac{R_i * C_j}{T}$$

where E_{ij} is the expected frequency for the cell in the ith row and the jth column, R_i is the total number of subjects in the ith row, C_j is the total number of subjects in the jth column, and T is the total number of subjects in the whole table.

3. Compute Chi Square using the formula:

$$\chi^2 = \sum_{ij} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

4. Compute the degrees of freedom using the formula:

df = (NR-1)(NC-1)

where NR is the Number of Rows and NC is the Number of Columns.

5. Use a chi square table to look up the probability value.
6. Determine the closest *p* (probability) value associated with your chi-square and degrees of freedom. If the *p* value for the

calculated χ^2 is p > 0.05, accept your hypothesis. If the p value for the calculated χ^2 is p < 0.05, reject your hypothesis.

2.7 Hypergeometric Distribution

The hypergeometric distribution arises when two sets are chosen from a larger set of elements. We want to test the hypothesis that the two sets were chosen at random and independently. Denote N the total number of objects, A the number of elements of the first set, B the number of elements of the second set, and t the number of elements in the intersection of the two sets. Let x be the random variable counting the size of the intersection, assuming the sets were chosen independently. Then the probability function F(x) is the hypergeometric distribution given by:

$$F(t) = Pr(x \leq t) = \sum_{i=0}^{t} \frac{\binom{A}{i}\binom{N-A}{B-i}}{\binom{N}{B}}$$

Thus, in order to give a p-value for over representation of the intersection, we need to compute:

$$Pr(x \geq t) = \sum_{i=t-1}^{\min(A,B)} \frac{\binom{A}{i}\binom{N-A}{B-i}}{\binom{N}{B}}$$

Hypergeometric Test for Three Sets

The above test is used to decide if two sets are chosen independently. This can be extended to a larger number of sets. For example, if we are interested in the dependence of choosing three sets from a larger set. We need to account here for the pairwise dependence between couples of sets. The null hypothesis will be that the choice of the three sets is independent given the pairwise dependencies. Let N be the size of the large set, and A, B and C the sizes of the three sets. Let AB, AC, and BC be the sizes of the pairwise intersections of the corresponding sets, and let t be the size of the intersection of the three sets. We assume that the three sets were chosen at random such that the pairwise intersections sizes are kept. Thus, if x is the random variable denoting the size of the three sets intersection, the distribution of x is given by:

$$Pr(x \leq t) = \frac{\sum_{i=m}^{t} \binom{AB}{i}\binom{B-AB}{BC-i}\binom{A-AB}{AC-i}\binom{N-A-B+AB}{C-AC-BC+i}}{\sum_{i=m}^{M} \binom{AB}{i}\binom{B-AB}{BC-i}\binom{A-AB}{AC-i}\binom{N-A-B+AB}{C-AC-BC+i}}$$

Where m and M are the minimal and maximal possible values of the intersections, given by:

$$M = min(AB, AC, BC),\ m = max(AB + AC - A, AB + BC - B, AC + BC - C, 0)$$

Again, in order to give a p-value for over representation of the intersection, we need to compute:

$$Pr(x \geq t) = 1 - Pr(x \leq t - 1)$$

2.8 Unsupervised Analysis

Super Paramagnetic Clustering (SPC) [10]

SPC is based on the physical properties of an inhomogeneous ferromagnetic. SPC uses a particular cost function for each partition and generates an ensemble of partitions at a fixed value of the average cost (average over the ensemble). The SPC cost function uses a distance function between the elements, and penalizes assignment of close elements to different partitions. The probability for a given partition configuration is given by the Boltzmann-Gibbs distribution where the temperature defines the average cost. At every temperature the probability that a pair of elements is assigned to the same partition is calculated, using an efficient Monte Carlo algorithm (cite Swendsen-Wang) by averaging sampled the different partition configurations at that temperature, according to their probabilities. Elements will be assigned to the same cluster only if they appear with a high enough probability in the same partition. Hence, for each temperature we have a different natural configuration of clusters. A stable cluster is a cluster that "lives" and does not separate into different groups for a large range ΔT.

The advantages of SPC are stability against noise, generating a hierarchy seen as a dendrogram ("tree view") and providing a way to recognize stable clusters, using a single distance function between the elements. In addition SPC does not need specification of the number of clusters in advance, a major advantage once working with large data sets, as microarray data. In particular, SPC provides a reliable stability index for clusters.

We used a new version of SPC (O. Barad, M.Sc thesis 2003), that uses mean field approximation instead of Monte Carlo in order to estimate the probability that a pair of elements is assigned to the same partition at a given temperature. The use of mean field approximation makes SPC deterministic, it reduces the running time of the probability estimation stage by a factor of 100 and the overall running time of SPC by factor of 10, and it has only minor effect of the clustering results. The new version enables us to cluster very large group of genes (~8000) and adjust the algorithm parameters.

Reference List

1. Thomson, J.A., et al., *Embryonic stem cell lines derived from human blastocysts.* Science, 1998. **282**(5391): p. 1145-7.
2. Amit, M., et al., *Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture.* Dev Biol, 2000. **227**(2): p. 271-8.
3. Dazard, J.E., et al., *Switch from p53 to MDM2 as differentiating human keratinocytes lose their proliferative potential and increase in cellular size.* Oncogene, 2000. **19**(33): p. 3693-705.
4. Jones, P.H. and F.M. Watt, *Separation of human epidermal stem cells from transit amplifying cells on the basis of differences in integrin function and expression.* Cell, 1993. **73**(4): p. 713-24.
5. Dazard, J.E., et al., *Genome-wide comparison of human keratinocyte and squamous cell carcinoma responses to UVB irradiation: implications for skin and epithelial cancer.* Oncogene, 2003. **22**(19): p. 2993-3006.
6. Sokal, R.R. and F.J. Rohlf, *Biometry: the principles and practice of statistics in biological research*. 3rd edition ed. 1995, New York: Freeman W. H. and Co. pp. 392.
7. Benjamini, Y.a.H., Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. Roy. Stat. Soc. B., 1995. **57**: p. 289–300.
8. Su, A.I., et al., *Large-scale analysis of the human and mouse transcriptomes.* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4465-70.
9. *http://faculty.vassar.edu/lowry/webtext.html*.
10. Blatt, M., S. Wiseman, and E. Domany, *Superparamagnetic clustering of data.* Physical Review Letters, 1996. **76**(18): p. 3251-3254.

Chapter 3

3 Results

3.1 Preprocessing and filtering

The expression data of the samples were organized in a matrix of *n_s=17* columns (experiments) and 22,283 rows (probe sets (PS) on the chip). We studied 17 samples: 3 ESC, 4 HSC and 4 HDC, 3 KSC and 3 KDC. 15427 PS with at least one “present” call, obtained from MAS 5.0, were selected, expression levels < 30 were thresholded to 30 and $\log_2$ was taken to generate the final gene expression matrix (17 x 15427). In the first analysis we filter the genes in the matrix using ANOVA [4]. We kept only the genes whose variance between groups is larger than the variance within each group. The p-value for this was calculated and false discovery rate (FDR) [5] was controlled at 0.001 to overcome the multiplicity problem. We have taken the ~5400 PS (4218 genes) that showed the largest inter-sample variation. The expression matrix of these PS, displayed in Fig 3.1, show that stem cell samples express many genes at a higher level than differentiated samples. This is the meaning of the zebra stripes seen in Fig 3.1. This observation suggests that in stem cells the genetic program primes the expression of a large number of genes which are downregulated or turned off upon differentiation. This scenario, of promiscuous gene expression in stem cells that prefaces the differentiated state, was already suggested in the case of hematopoietic stem cell differentiation on the basis of expression of erythroid or granulocyte markers in the progenitor cell prior to commitment [1], and recent work extended this observation also to the analysis of genes in the hematopoietic system [2, 3].

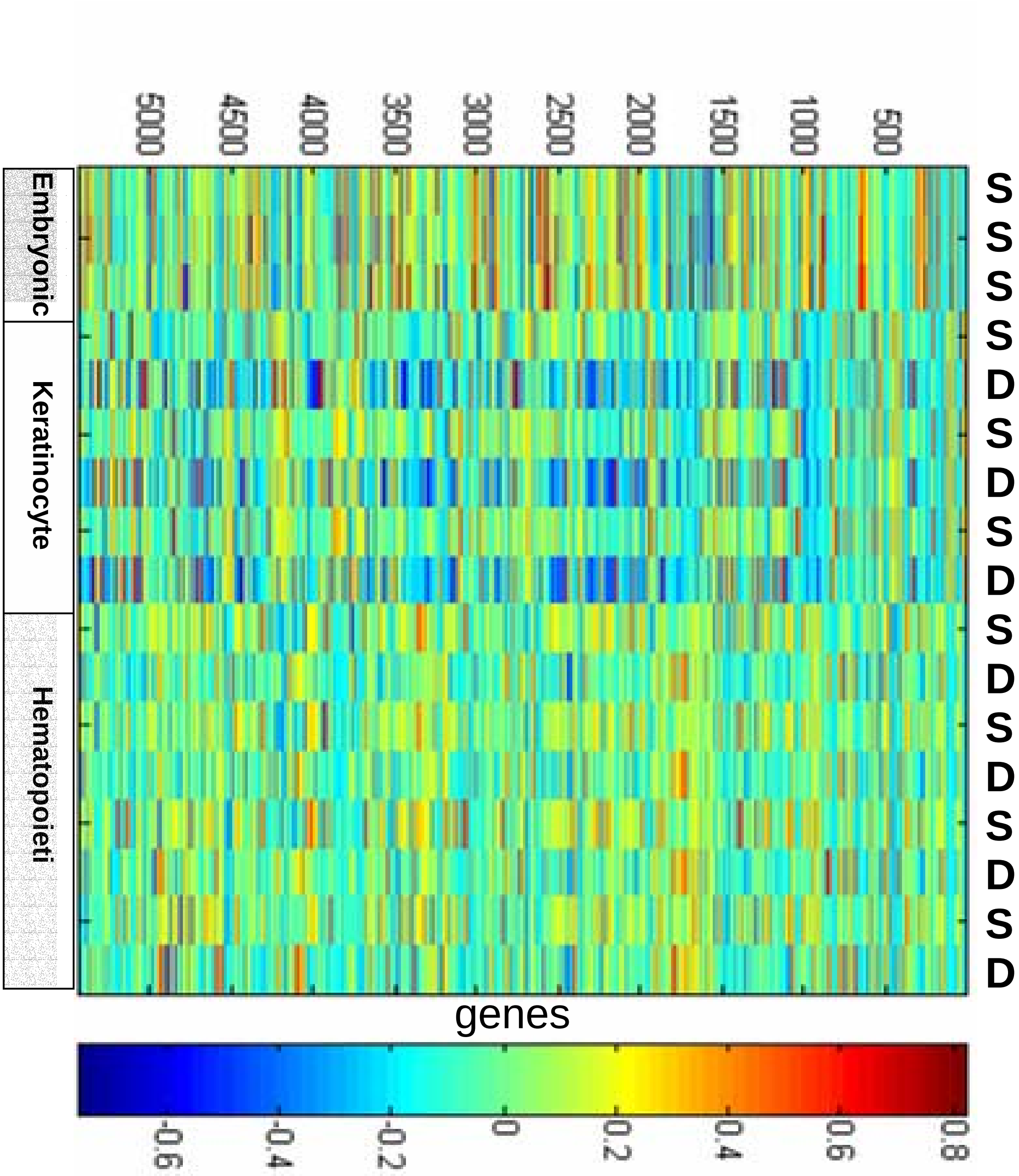


Figure 3.1 Centered and normalized expression level of ~2600 probe sets (PS) which showed the largest inter sample variation. "S" denotes a stem cell sample "D" denotes a differentiated sample. A pronounced pattern shows high expression values in the stem cell samples versus the differentiated ones. The overall expression level of all chips was scaled to the same value; the pronounced difference was seen when we looked at the genes with highest variation.

We wanted to check the difference between the expression levels of all genes in stem cell samples vs. differentiated cell samples. To this end, we examined the distribution of expression in stem cell samples vs. the mature ones (see Fig 3.2). These distributions have already been scaled according to Affymetrix scaling factors within acceptable limits (see chapter 2 – Materials and Methods). Therefore, the trimmed mean intensities (cut 2% low and high outliers) for all arrays in the experiment were equal. Nevertheless, the number of probe sets which have signal values between 100-500 is higher in stem cell samples vs. differentiated ones (see fig 3.2).

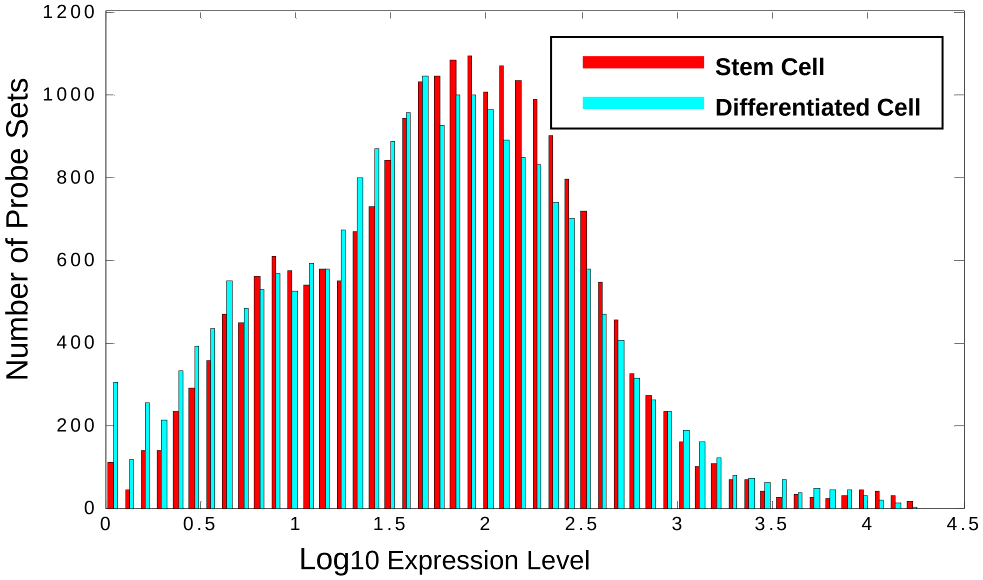


Figure 3.2 Keratinocyte stem cell sample distribution vs. keratinocyte differentiated cell sample. The stem cell sample, marked by red histograms, has a higher number of probe sets between expression levels 10^2 to $10^{2.7}$ than the mature cell sample, marked by the light blue histogram. Scaling factors were carried out for all arrays within acceptable limits (0.86-1.26 fold). Therefore, the trim mean intensities (cut 2% low and high outliers) for all arrays in the experiment are equal.

Considering the observed finding that stem cells express large numbers of genes which are downregulated or turned off upon differentiation, we decided to divide the samples into two differentiation pathway groups. By doing this, we tested the changes of gene expression during differentiation.

(A) Hematopoietic (H) pathway; ESC → HSC → HDC (3+4+4 samples), and

(B) Keratinocytic (K) pathway; ESC → KSC → KDC (3+3+3 samples).

For each group the genes were filtered using ANOVA [4]. We kept only the genes whose the variance of between groups is larger than the variance within each group. The p-value for this was calculated and false discovery rate (FDR) [5] was controlled at 0.05 to overcome the multiplicity problem. This left 8290 PS (6293 genes) that vary significantly over the three kinds of cell states in group (A) and 5432 (4301 genes) for group (B); the reason for this difference was the different numbers of samples in the two groups.

3.2 Stem Cells Expressed Thousands of Genes that are Markedly down Regulated upon Differentiation

We present in Figs. 1a and 1b the expression levels of the significantly varying PS. The data shows that ESC (black line, Fig. 3.3) express many genes at a higher level than any other cell and the majority of transcripts exhibit marked down regulation along the differentiation pathway. 4392 PS are down regulated as cells differentiate from ESC to HSC (green dots, Fig. 3.3a), followed by a further downward shift upon progression from each HSC to its differentiated counterpart (red dots, Fig. 3.1a). In contrast, this is accompanied by up-regulation of a smaller group of 2638 PS, with low expression in ESC and high in the HDC. A similar pattern is seen in the keratinocytic pathway (Fig. 3.3b). 3417 PS are down regulated as cells differentiate from ESC to KSC (green dots, Fig. 3.1b), followed by a further downward shift upon progression from each KSC to its differentiated counterpart (red dots, Fig. 3.3a). In contrast, this is accompanied by up-regulation of a smaller group of 1423 PS, with low expression in ESC and high in the KDC.

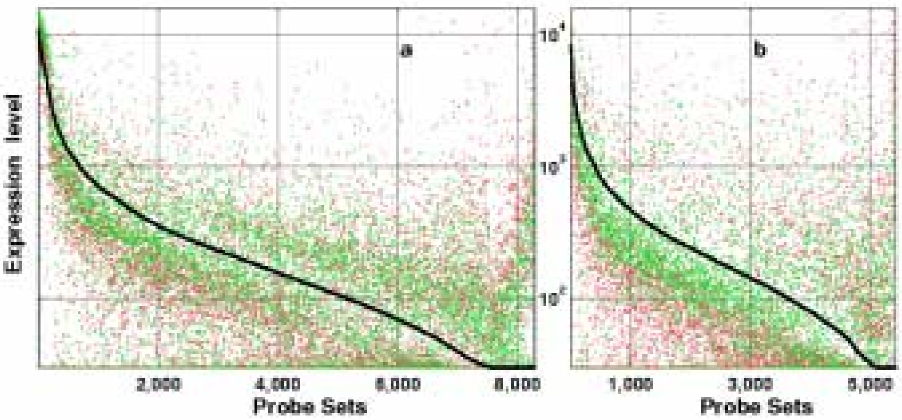


Figure 3.3 Expression levels of probe-sets (PS) that vary significantly between ESC, ASC and differentiated cells. The PS were sorted according to their ESC expression levels, marked by black circles (that form a line). The expression levels in HSC or KSC are indicated by green dots and in HDC and KDC by red dots. **a** Expression levels of 8290 PS that vary between ESC, HSC and HDC. **b** Expression of 5432 PS that vary between ESC, KSC and KDC. This difference in PS numbers between (A) and (B) is due to the different numbers of samples [4]. Only PS with P-values that passed ANOVA at an FDR level of 0.05 were plotted. In **a** 4683 PS are expressed at lower level for HDC vs. ESC while 3562 PS expressed at higher level for HDC vs. ESC. In **b** these numbers are 3626 and 1791 respectively.

We looked for an underlying design principle that could explain these results. A prime candidate for pluripotential differentiation is the parsimonious "just in time" strategy; expressing genes only when needed, i.e. at the moment of commitment to a particular differentiation path. The opposite extreme is the seemingly more wasteful "just in case" strategy, which keeps a wide repertoire of expressed genes, to be present in case a particular path is selected. We will address this question further on.

3.3 Clustering Analysis Shows Distinct Self-renewal Genes for Different stem Cell Tissues

We clustered [6] the samples of groups (A) and (B) separately, to identify distinct differentiation-induced variations of the expression profiles, and to assign genes to clusters of similar patterns of expression. Fig. 3.4 and 3.5 depicts the expression matrix after clustering, centered and normalized of the genes in the H (Fig. 3.4) and K (Fig. 3.5) pathways. Six clusters are clearly shown. Clusters 1, 2 and 3 contain ESC genes that are down-regulated with differentiation in both H (H1-H3) and K (K1-K3) pathways. Clusters 4 and 5 contain genes that are upregulated along the differentiation pathway and clusters 6 contain genes expressed only in adult stem cells (ASC). Clearly, ESC and ASC have different gene expression profiles.

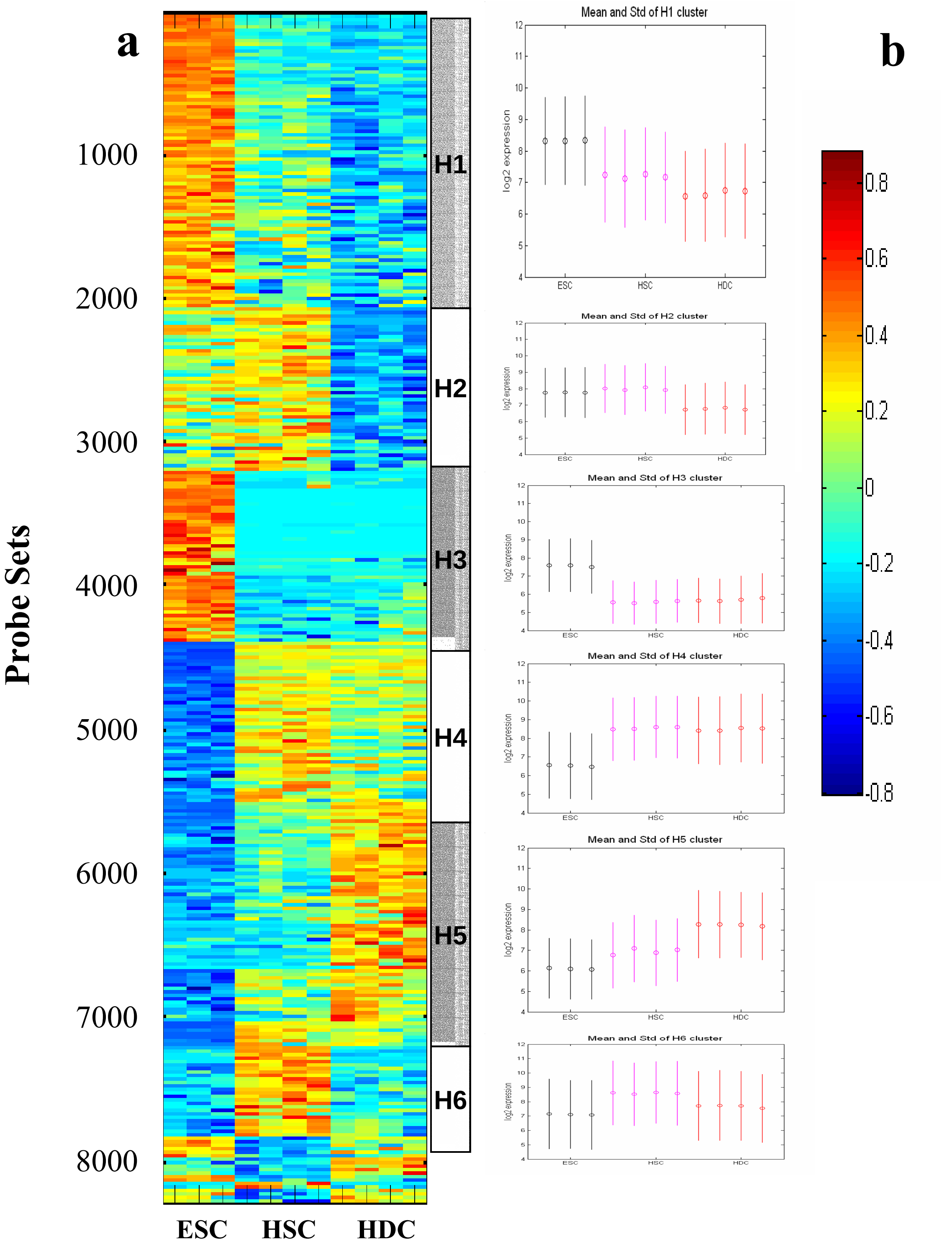


Figure 3.4 Clustering analysis of PS expression levels in hematopoietic pathways. The expression levels of the PS taken from Fig. 3.3a were centered and normalized and the PS were reordered according to the dendrogram produced by the SPC algorithm [6]. **a** Expression matrix of 8290 PS in ESC, HSC, and HDC. **b** Corresponding expression profiles of the raw data of each cluster [mean +/- std].

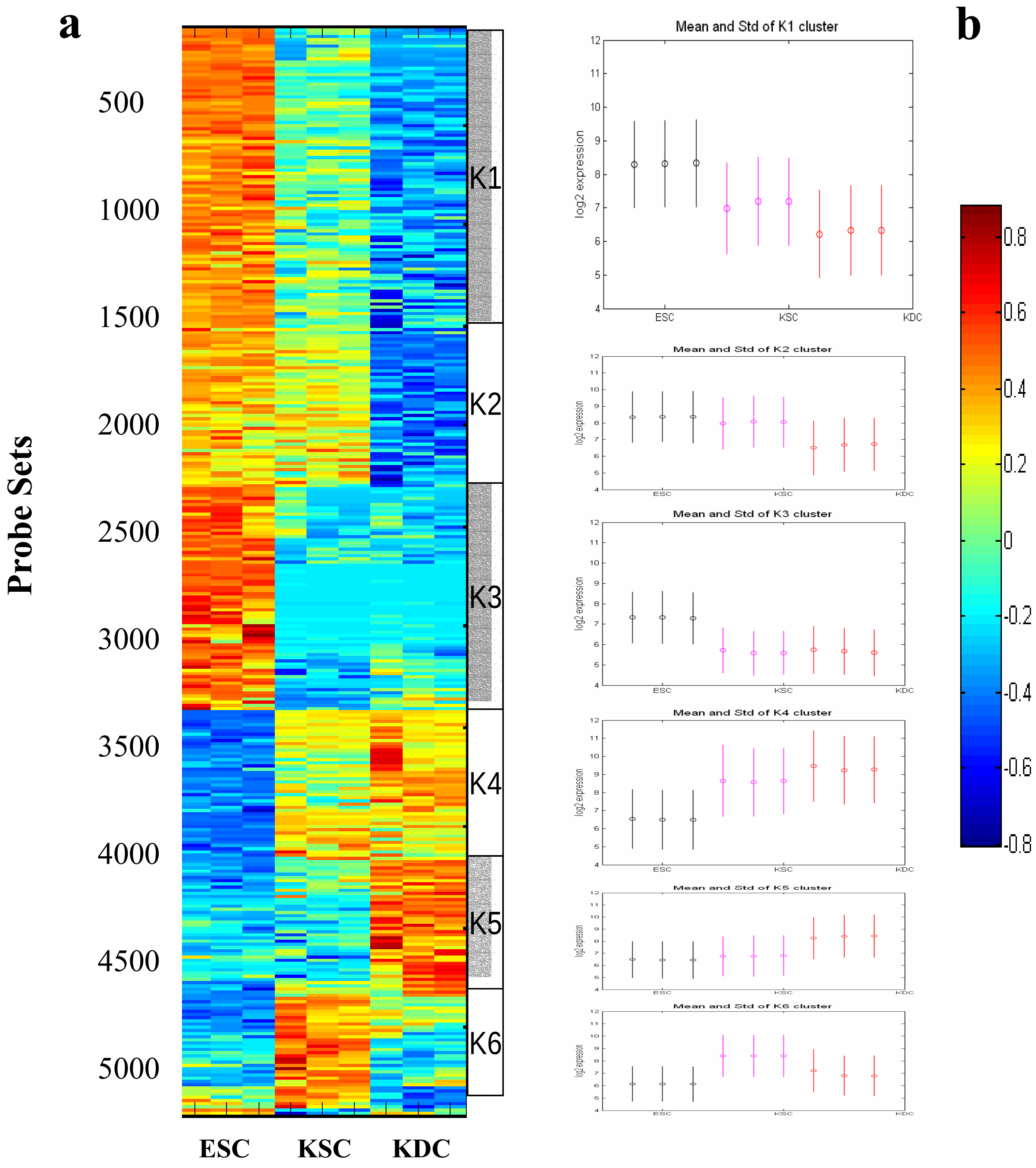


Figure 3.5 Clustering analysis of PS expression levels in keratinocytic pathways. The expression levels of the PS taken from Fig. 3.3b were centered and normalized and the PS were reordered according to the dendrogram produced by the SPC algorithm [6]. **a** Expression matrix of 5432 PS in ESC, KSC, and KDC after centered and normalized **b** Corresponding expression profiles of the raw data of each cluster [mean +/- std].

We looked in the literature for the known biological function of all stem cell genes in our clusters. Table 3.1 presents the function of selected genes that were previously shown to be typical of one of the cell stages, and belong to one of the six clusters shown in Fig. 3.4 and 3.5. Clusters 1 and 2 contain genes that are common to ESC and ASC and therefore may represent the “stemness” genes as previously defined [7, 8]. It should be noted, however, that many genes, well known to be markers for undifferentiated ESC or related to ESC self-renewal, belong to clusters H3 (Fig. 3.4) and K3 (Fig. 3.5), and thus are suppressed in ASC. For example, NANOG is known to be capable of maintaining ESC self-renewal. Experiments on nanog-deficient cells failed to generate epiblast and produced only parietal endoderm-like cells. These cells lost pluripotency and differentiated into extra-embryonic endoderm lineage. Other examples of genes associated with ESC self-renewal or known to be markers for ESC are POU5F1 (OCT4), SOX2, FOXH1, TDGF1 (Cripto), LeftyA & B, Thy1 [9-13] – see Table 3.1. Hence, these genes are not responsible for self-renewal in ASC. Their roles are apparently taken over in ASC by genes of clusters H6 or K6, which show expression only in ASC (neither in ESC, nor in mature cells), and indeed contain genes known to be essential for the self-renewal of ASC, progenitors and tissue development (e.g. TP73L (p63), ITGB4 and BNC for skin [14-16], and e.g. BMI1, CD34, TIE, KIT, TAL1 (SCL), and RUNX1 for blood [17-19] – Table 3.1). These observations indicate that the common genes in ESC and ASC cannot define the so-called “stemness” genes. Rather, there are two distinct groups of genes characteristic of stem cells: those common to ESC and ASC (from clusters 1+2) and those specific of each kind of SC (from cluster 3 for ESC, cluster H6 for HSC, and cluster K6 for KSC).

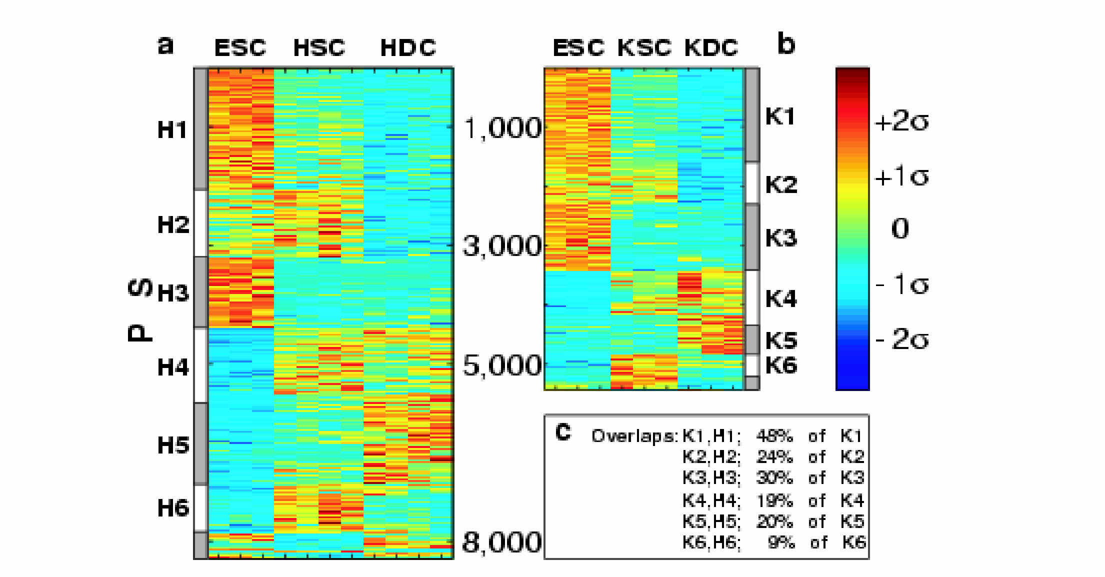


Figure 3.6 Clustering analysis of PS expression levels in hematopoietic and keratinocytic pathways. The expression levels from Fig. 3.4 and 3.5. **a** Expression matrix of 8290 PS in ESC, HSC, and HDC. **b** Expression matrix of 5432 PS in ESC, KSC and KDC. **c** Percentages of overlaps between the related 6 clusters were calculated relatively to keratinocyte clusters.

Table 3.1 Selected genes identified in clusters of Fig. 3.2 that are known to be important in the various cell stages: ESC, HSC, KSC, HDC and KDC.

	Hematopoietic clusters				**Keratinocytic clusters**		
	Identifier	**Symbol**	**Short Name**		**Identifier**	**Symbol**	**Short Name**
H1	X52078.1	TCF3	transcription factor 3	**K1**	BG393795	TCF3	transcription factor 3
	BF510715	FGF4	FGF4		BF510715	FGF4	FGF4
	NM_014366.1	NS	Nucleostemin		U91903.1	FRZB	frizzled-relat. prot.
	L37882.1	FZD2	frizzled 2		NM_001845.1	COL4A1	collagen, type IV, α1
	NM_000435.1	NOTCH3	Notch 3		AK026737.1	FN1	fibronectin 1
	AF029778	JAG2	jagged 2		NM_021953.1	FOXM1	forkhead box M1
	AL556409	GAL	Galanin		AL556409	GAL	galanin
	NM_005842.1	SPRY2	sprouty 2		NM_005359.1	MADH4	SMAD4
H2	AK026674.1	TCF4	transcription factor 4	**K2**	BC004912.1	BPAG1	bullous pemph. ag.1
	NM_001331.1	CTNND1	δ catenin 1		NM_003798.1	CTNNAL1	α catenin like 1
	M87771.1	FGFR2	KGF receptor		NM_022969.1	FGFR2	KGF receptor
	NM_006017.1	CD133	prominin-like 1		NM_003012.2	SFRP1	frizzled-relat. prot. 1
	NM_003506.1	FZD6	frizzled homolog 6		NM_000165.2	GJA1	connexin 43
	NM_000165.2	GJA1	connexin 43				
	NM_005631.1	SMO	Smoothened				
	NM_003107.1	SOX4	SRY-box 4				
H3	AF268613.1	POU5F1	OCT4	**K3**	AF268613.1	POU5F1	OCT4
	NM_024674.1	LIN-28	RNA-binding protein		NM_024674.1	LIN-28	RNA-binding protein
	NM_003212.1	TDGF1	Cripto		NM_003212.1	TDGF1	Cripto
	NM_024865.1	NANOG	ES transcription factor		NM_024865.1	NANOG	ES transcription factor
	NM_003240.1	EBAF	left-right determ. fact. A		NM_003240.1	EBAF	left-right determ. fact. A

	NM_020997.1	LEFTB	left-right determ. fact. B		NM_020997.1	LEFTB	left-right determ. fact. B
	NM_003577.1	UTF1	ES transcription factor 1		NM_003577.1	UTF1	ES transcription factor 1
	AA218868	THY1	Thy-1 cell surface ag.		AA218868	THY1	Thy-1 cell surface ag.
	NM_001290.1	LDB2	LIM domain binding 2		NM_001290.1	LDB2	LIM domain binding 2
	NM_003923.1	FOXH1	forkhead box H1		NM_003923.1	FOXH1	forkhead box H1
	AF202063.1	FGFR4	FGFR4		NM_002011.2	FGFR4	FGFR4
	L07335.1	SOX2	SRY-box 2		L07335.1	SOX2	SRY-box 2
	NM_016941.1	DLL3	delta-like 3		NM_016941.1	DLL3	delta-like 3
	NM_005585.1	MADH6	SMAD6		NM_005585.1	MADH6	SMAD6
	NM_001134.1	AFP	alpha-fetoprotein		NM_001134.1	AFP	alpha-fetoprotein
	NM_007295.1	BRCA1	breast cancer 1		AF005068.1	BRCA1	breast cancer 1
	U96136.1	CTNND2	δ catenin 2		AF035302.1	CTNND2	δ catenin 2
	NM_017412.1	FZD3	frizzled homolog 3		NM_017412.1	FZD3	frizzled homolog 3
	NM_020634.1	GDF3	growth diff. factor 3		NM_020634.1	GDF3	growth diff. factor 3
	U91903.1	FRZB	frizzled-related protein		NM_001463.1	FRZB	frizzled-related protein
	NM_012259.1	HEY2	hairy/enh. of split 2 YPRW		AF098951.2	ABCG2	ATP-bind. cassette G2
	U43148.1	PTCH	Patched				
	NM_003392.1	WNT5A	development regulator				
H4	NM_014676.1	PUM1	pumilio 1	**K4**	NM_005620.1	S100A11	calgizzarin
	D87078.2	PUM2	pumilio 2		NM_002965.2	S100A9	calgranulin B
	BC005912.1	FCER1A	Fc frag. of IgE, high aff. I		NM_002966.1	S100A10	calpactin I
	NM_019102.1	HOXA5	homeo box A5		NM_005978.2	S100A2	CAN19
	BC005332.1	IGKC	Ig const. κ		NM_003125.1	SPRR1B	cornifin
	BG340548	IGHM	Ig heavy const. m		NM_002203.2	ITGA2	integrin α2
	NM_005574.2	LMO2	LIM domain only 2		NM_005547.1	IVL	involucrin
	AA573862	HLA-A	MHC I, A		NM_005046.1	KLK7	kallikrein 7
	X76775	HLA-DMA	MHC II, DM β		M19156.1	KRT10	keratin 10
					X57348	SFN	stratifin
H5	NM_001738.1	CA1	carbonic anhydrase I	**K5**	AL356504	FLG	filaggrin
	NM_000129.2	F13A1	coag. factor XIII, A1		AF243527	KLK5	kallikrein 5
	U62027.1	C3AR1	compl. comp. C3a R1		NM_006121.1	KRT1	keratin 1
	AF130113.1	CYB5-M	cytochrome b5 prec.		NM_002274.1	KRT13	keratin 13
	NM_001978.1	EPB49	eryth. memb. prot. 4.9		NM_000427.1	LRN	loricrin
	NM_004107.1	FCGRT	Fc frag. of IgG, α		NM_002963.2	S100A7	psoriasin 1
	NM_005143.1	HP	Haptoglobin		NM_003238.1	TGFB2	TGFβ2
	NM_000558.2	HBA1	hemoglobin α1		NM_004245.1	TGM5	transglutaminase 5
	H53689	IGL@	Ig l locus				
	BE138825	HLA-F	MHC I, F				
	NM_002120.1	HLA-DOB	MHC II, DO β				
H6	M81104.1	CD34	CD34 antigen	**K6**	NM_001717.1	BNC	basonuclin
	NM_005180.1	BMI1	B lymph. MLV ins. reg.		AF091627.1	TP73L	p63
	NM_000222.1	KIT	SCF receptor		NM_002204.1	ITGA3	integrin α3
	NM_005424.1	TIE	endothelial RTK		NM_000213.1	ITGB4	integrin β4
	NM_003189.1	TAL1	SCL		NM_001723.1	BPAG1	bullous pemph. ag.1
	D43968.1	RUNX1	RUNT TF 1		NM_000494.1	BPAG2	collagen XVII □1
	AL134303	EGFL3	EGF-like-domain 3		NM_000227.1	LAMA3	laminin α3
	NM_018951.1	HOXA10	homeo box A10				

3.4 Programming Pluripotency of Stem Cells Involves Genes Used by Many Tissues

The clustering results show that when going from ESC to adult differentiated cells, in the hematopoietic pathway 4392 PS (3483 genes) are down regulated and 2638 PS (1998 genes) are upregulated, while in the keratinocyte pathway 3417 PS (2758 genes) are down regulated and 1423 PS (1115 genes) are upregulated. The massive down regulation is consistent with the "just in case" design principle underlying pluripotential differentiation. Our data suggest that in order to maintain their potential for pluripotency, ESC "keep their options open" by promiscuous gene expression, maintaining thousands of genes at intermediate levels, to be down-regulated upon commitment to a cell fate for which they are not needed. This down-regulation is required for establishing the differentiated state. The strategy is apparently universal; it holds for differentiation from ESC to adult SC and also for passage from the latter to mature tissue. It also holds irrespective of the particular differentiation pathway (H or K). Our interpretation for the connection between changes in genes expression and differentiation is supported by the fact that among the genes of clusters H1, H2 and H3, or K1, K2 and K3, many are high in ESC and low or absent in the adult tissue (HDC or KDC); hence they are not needed to produce these tissues. We hypothesize that most of the multitude of transcripts, which are down regulated upon differentiation towards a tissue A, represent other optional cell fates, and may be needed by the ESC to produce other tissues e.g. B, C, D. In parallel to down regulation of many genes, we observed a fairly large group (clusters 4 and 5), that are mostly low in ESC and upregulated upon terminal differentiation. These genes are needed mainly to produce the target tissue A (keratinocytic or hematopoietic in our case). Indeed, clusters K4+K5 contain a large fraction of skin-specific genes (e.g. keratins, kallikreins, cornifin, involucrin and filaggrin) and in H4+H5 we find blood specific genes like hemoglobin, immunoglobulin chains, histocompatibility genes and others (Table 3.1). Our hypothesis is that an embryonic stem cell expresses genes that are used in adult tissues. In other words, the multipotential embryonic stem and progenitor cells prime several different lineage-affiliated programs of gene activity prior to unilineage commitment and differentiation. To check this hypothesis, we have looked in the literature for an experiment that includes a wide group of adult tissues in humans. The only experiment we found was the dataset of Su et al [20] on 20 tissues (GNF dataset at http://expression.gnf.org/cgi-bin/index.cgi), which we supplemented by gene expression measurements in 4 normal human epidermal keratinocyte samples [21], that used the same U95 Affymetrix chip.

Su et al [20] have generated and analyzed gene expression from a set of samples spanning a broad range of biological conditions. Specifically, they profiled gene expression from 91 human and mouse samples across a diverse array of tissues, organs, and cell lines. Because these samples predominantly came from the normal physiological state in the human and mouse, this dataset represents a preliminary, but substantial, description of the normal mammalian transcriptome. Su et al [20] have identified tissue specific genes according to the following conservatively defined filtering criteria: a tissue-specific gene must have an expression level greater than 200 in one tissue, and less than 100 in all other tissues. This analysis, performed for all tissues in both mouse and human datasets, identified 311 human and 155 mouse tissue specific genes with known function, and 76 human and 101 mouse genes whose functions were previously uncharacterized.

We refer as GNF* to those of our extended GNF genes that appear also on the U133 Affymetrix chip. First, we identified in GNF* those genes that appeared in one of the clusters H1-H6 or K1-K6. Each such gene was tested for tissue specificity, but using a different criterion: We defined a gene as specific if it is highly expressed in one tissue versus all the others (see Z-Score analysis in Materials and Methods – chapter 2).

Let us now predict what the results would be if ESC indeed use the "just in case" strategy. Say genes are expressed, for the eventuality that they become needed (in case of commitment to a yet unknown fate). Then it makes sense to express preferentially genes that are needed by several tissues. Hence, we would expect to see that ESC preferentially express genes which are expressed also in many adult tissues, just in case they will be needed upon commitment into a cell fate. Upon commitment to a cell type which does not need such a gene, its expression will shut down. Hence such genes are expected to be found in clusters H1-H3 (or K1-K3). In contrast, in clusters H4+H5 (or K4+K5) we would expect to find blood (keratinocyte) specific genes only.

And indeed we found, in agreement with our model, that clusters H4+H5 of Fig. 3.4 contain 370 genes specific to blood and related tissues like spleen and thymus. Clusters K4+K5 contain skin specific genes. On the other hand, none of the clusters H1, H2, H3 contain significant numbers of blood specific genes (Fig. 3.7a). Also, none of the clusters K1, K2, K3 contain significant numbers of skin specific genes (Fig 3.7b).

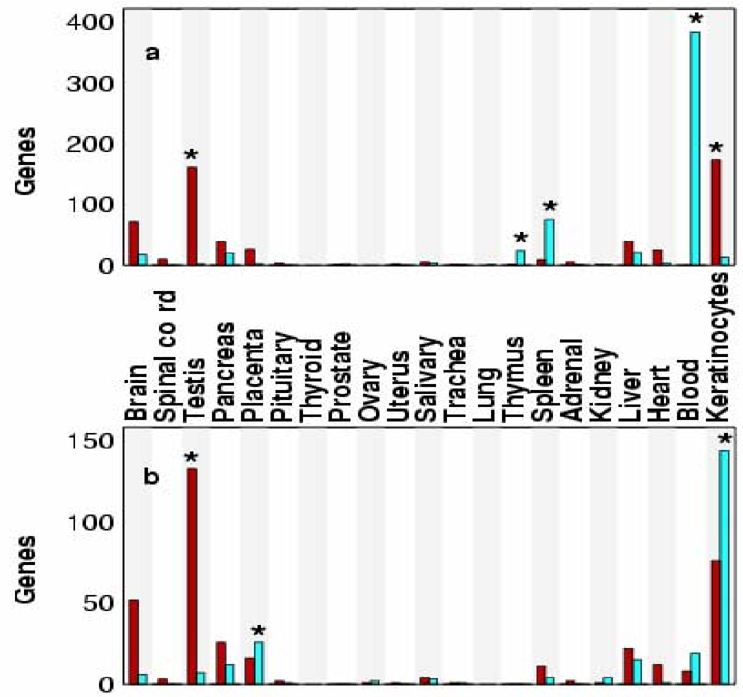


Figure 3.7 Distribution of Tissue-Specific Genes in Hematopoietic and Keratinocytic Pathways. Tissue specific genes (see Z-score analysis in Materials and Methods) obtained from supplemented GNF* dataset were determined for the genes in the clusters of Fig. 3.4 and 3.5. Tissue-specific genes corresponding to clusters 4 and 5 of Fig 3.4 are represented by blue bars and tissue-specific genes corresponding to clusters 1, 2 and 3 are represented by red bars. **a.** Tissue specific genes in hematopoietic pathway. **b.** tissue specific genes in keratinocytic pathway. Statistically significant numbers of tissue specific genes are labeled with a star.

Recall that Su at al found only a few hundred tissue-specific genes - and our definition yields also about the same number. Hence limiting our attention to tissue specific genes restricted our analysis to only these few hundreds of genes. As indeed can be seen from Fig. 3.7, cluster 1, 2 and 3 do not contain significant numbers of tissue specific genes (except for the testis). Since we wanted to understand the roles of the thousands of genes that were down regulated through differentiation in Fig 3.4 and 3.5 and were not specific for any tissue, we divided the genes of GNF* according to the number of tissues in which their expression [20] is high (exceeds 500). If the number of such tissues is 1 – 4, we termed the gene “tissue affiliated”, and if it was greater than 4, the gene was termed “expressed in many tissues”. In clusters H4+H5, we found mainly “tissue affiliated” genes needed for blood (striped in Fig. 3.8a), and related tissues like spleen and thymus (Fig. 3.8a). A gene that was high in some other tissue (e.g. pancreas) was most likely to be “expressed in many tissues” in addition to having a high expression level in blood. On the other hand, clusters H1+H2+H3 contained mostly genes that were “expressed in many tissues” (yellow in Fig. 3.8b), and a smaller number of "tissue affiliated" genes, but not blood specific ones. The genes of H1, H2, and H3 were expressed at a relatively high level in ESC ("just in case" they are needed) and since they were not needed in blood, they were turned off upon commitment to blood. Similar analysis is presented for K pathway (Fig. 3.9). These conclusions were also supported by a χ^2 test (Table 3.2). These results support one of the main properties of stem cells. Embryonic stem cells do not have any specific structure (like a mature cell) that would allow them to perform specific functions. Therefore, an ESC can not express many "tissue affiliated" genes i.e. genes expressed only in cells of a particular tissue, which enable the specific structure of the cell or its function. However, an unspecialized ESC can give rise to all specialized cell types. For this reason, it keeps a small number of "candidate" specific genes, that perhaps play a role in triggering the differentiation process upon commitment. Turning back to the question of which strategy stem cells use for pluripotency, our results indicate that they follow both strategies: stem cells keep thousands of non specific genes expressed (“just in case”); most of these are quenched upon differentiation, if not needed. However, target tissue specific genes are up-regulated when needed (“just in time”) to determine the cell fate. These finding are relevant to the question of pluripotency and plasticity of adult stem cells.

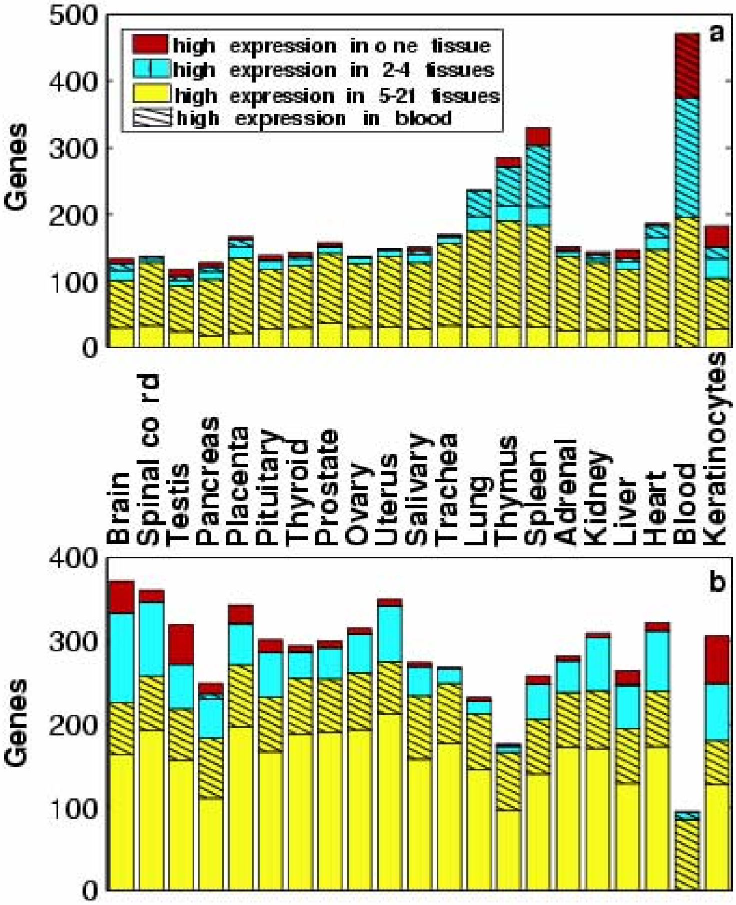


Figure 3.8 Distributions in 21 Tissues of Genes that Change Expression in the Hematopoietic Pathway. We used GNF*, the supplemented GNF dataset, to identify genes from the clusters of Fig. 3.4 and 3.5, that have high expression (>500) in various differentiated tissues [20]. **a** The distribution of those genes of clusters H4, H5 that have low expression (< 200) in ESC. **b** The distribution of those genes of clusters H1, H2, and H3 for which expression in HDC < 200. Colors indicate the number of tissues in which a gene is highly expressed and stripes indicate that the gene is high in blood. Note that about 80 genes had high expression in blood in Su et al. [20] and low in our data.

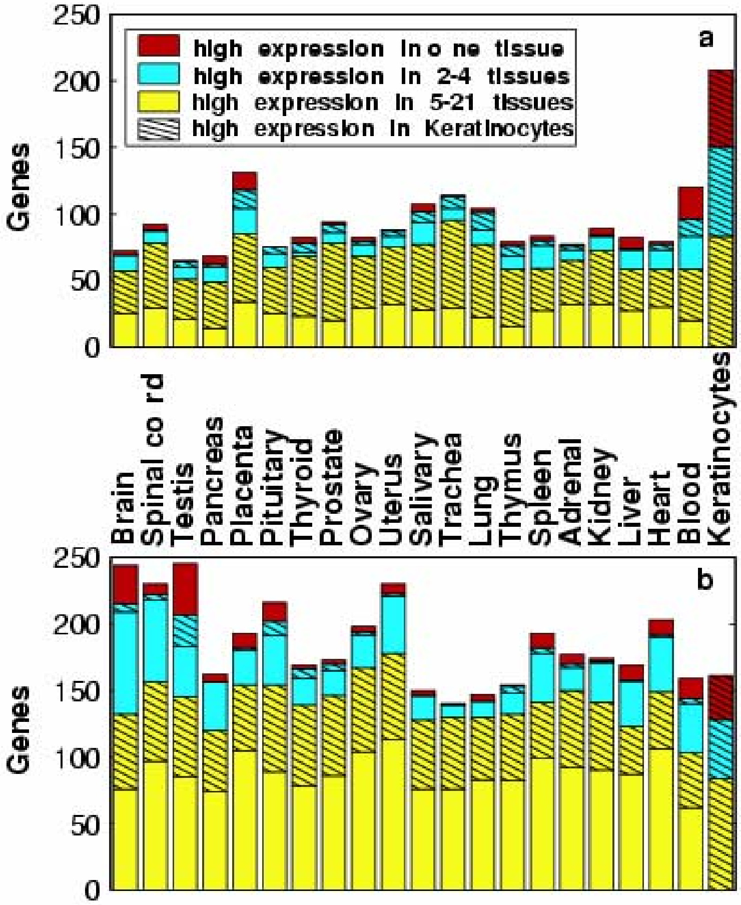


Figure 3.9 Distributions in 21 Tissues of Genes that Change Expression in the keratinocytic Pathway. We used GNF*, the supplemented GNF dataset to identify genes from the clusters of Fig. 3.4 and 3.5, that have high expression (>500) in various differentiated tissues [20]. **a** The distribution of those genes of clusters K4, K5 that have low expression (< 200) in ESC. **b** The distribution of those genes of clusters K1, K2, and K3 for which expression in KDC < 200. Colors indicate the number of tissues in which a gene is highly expressed and stripes indicate that the gene is high [20] in keratinocyte.

Table 3.2 Chi-Square test analysis indicates the relationship between "tissue-affiliated" genes and expression levels in ESC.

	Number of genes with high (>500) expression in	
	1– 4 tissues	**more than 4 tissues**
low (≤200) expression in ESC	1695	736
high (>200) expression in ESC	863	876

This analysis is based on expression in adult tissues [20] and may not reflect developmental potential. Hence we searched for tissue specific developmental genes for comparison with ESC and ASC, and used the data of pancreas specific genes reported by Wells [22]. Wells investigated the genes expressed in embryonic rat pancreas. He divided the embryonic pancreas to cells from the endocrine part and cells from the exocrine part. His results were summarized in three tables of genes: one was the endocrine pancreas development genes, the second was exocrine pancreas development genes and the third contained genes related to adult pancreas genes. Table 3.2 depicts the presence of developmental pancreas specific genes in the ESC clusters shown in Fig. 3.4 and 3.5. A large proportion of genes specific for pancreas development were expressed in ESC (31% in H pathway and 23% in K pathway; most of them were in clusters 1-3, Table 3.3). This provides further support to the notion that the genetic program of ESC contains many transcripts needed for a variety of tissues (in this case pancreas) that will be shut down upon cell fate commitment or conversely upregulated if differentiation signals induce one of these corresponding tissues. This co-expression of multitudes of genes at the ESC stage may be the basis for their pluripotency and provide options for future diverse cell fates. Hence stem cells express a large repertoire of genes and then select a few for continued expression as they differentiate to a target tissue. In people who suffer from type 1 diabetes, the cells of the pancreas that normally produce insulin are destroyed by the patient’s own immune system. Understanding the nature of the developmental genes expressed in embryonic stem cells may indicate the possibility to direct the differentiation of human embryonic stem cells in cell culture to form insulin-producing cells that eventually could be used in transplantation therapy for diabetics.

Table 3.3. Developmental pancreas specific genes in clusters of Fig. 3.4 and 3.5

	Hematopoietic clusters				Keratinocytic clusters		
	Identifier	**Symbol**	**Name**		**Identifier**	**Symbol**	**Name**
H1	NM_001954.2	DDR1	discoidin receptor	**K1**	NM_000088.1	COL1A1	Collagen,
	L37882.1	FZD2	frizzled		NM_006195.1	PBX3	pre-B-cell TF
	NM_012193.1	FZD4	frizzled		NM_003477.1	PDX1	Pancreatic homeobox
	NM_006870.2	DSTN	destrin		NM_002293.2	LAMC1	Laminin
	NM_002293.2	LAMC1	laminin				
	NM_000435.1	NOTCH3	Notch				
	AL157414	BMP7					
	NM_023107.1	FGFR1					
	U15979.1	DLK1	delta-like 1				
	NM_003477.1	PDX1	Pancreatic homeobox				
	NM_006195.1	PBX3	pre-B-cell TF				
	NM_000484.1	APP	amyloid β				
H2				**K2**	NM_003012.2	SFRP1	Frizzled
					NM_022969.1	FGFR2	
H3	NM_022969.1	FGFR2		**K3**	NM_004305.1	BIN1	Bridging integrator 1
	NM_000142.2	FGFR3			NM_012193.1	FZD4	frizzled
	NM_002011.2	FGFR4			NM_002011.2	FGFR4	
	NM_003012.2	SFRP1	secreted frizzled		AB028641.1	SOX11	SRY box 11
	AB028641.1	SOX11	SRY		L37882.1	FZD2	Frizzled
	NM_000638.1	VTN	vitronectin		NM_000638.1	VTN	vitronectin
	NM_000088.1	COL1A1	collagen		M25915.1	CLU	clusterin
	AF039555.1	VSNL1	visinin-like		AB028973.1	MYT1	myelin TF 1
H4	NM_004305.1	BIN1	bridging integrator 1	**K4**	NM_000700.1	ANXA1	Annexin A1
	NM_000700.1	ANXA1	annexin A1		NM_001305.1	CLDN4	claudin 4
	NM_001913.1	CUTL1	cut-like		NM_006870.2	DSTN	destrin
H5	AA809056	ACTB	actin β	**K5**			
	NM_002087.1	GRN	granulin				
H6				**K6**	AV733308	ITGA6	integrin α6
					NM_003385.1	VSNL1	visinin-like 1

3.5 Searching for "stemness" genes

We found, using clustering analysis, that there are different genes related to self renewal in each stem cell type. Nevertheless, we tried to repeat the work that was done previously [7] [8] (see chapter 1 Introduction) in order to search for new “stemness” genes in our stem cells. i.e. self-renewal and pluripotency genes shared by all stem cells. We selected genes that showed enrichment in KSC (and HSC) by at least 2-fold change of expression compared with their terminally differentiated counterparts KDC (or HDC). ESC enriched genes were selected (2-fold change) over KDC and HDC separately, and then by intersecting unigene accession numbers. As shown in the Venn diagram (Fig. 3.10), the intersection of the three lists of genes enriched in each individual SC, as determined by fold change analysis, contains 317 candidate stemness genes, enriched in all three stem cells. The probability of observing an overlap by chance as estimate using hyper geometrical distribution (see Materials and Methods –Chapter 2) is p = 0.9885. The hyper geometrical test provides an estimate to the statistical significance of the overlap of the three lists, for the given values of the pairwise overlaps between the three lists. The high p-value indicates that this number, of 317 genes, is expected to be obtained by chance! On the other hand, when we computed the probability of observing the overlap based on genes commonly expressed in two types of stem cells (two lists only), the probability drops dramatically (p-value is $p=10^{-235}$).

We then intersected these 317 genes found by fold-change analysis with those found by two statistical tests: Wilcoxon rank sum test and t-test. We selected genes that showed differential expression between two groups of samples: KSC (and HSC) compared with their terminally differentiated counterparts KDC (or HDC). ESC genes were selected over KDC and HDC separately, and then by intersecting unigene accession numbers (FDR controlled to all theses tests). These two statistical tests revealed a core of 271 genes that were contained in the 317 genes. Hence the genes found in the three-way intersection do not depend significantly on whether one uses fold-change or standard statistical tests to identify them. Out of these 271 genes, 263 (97%) and 235 (87%) belong to clusters H1+H2 and K1+K2 respectively (See Appendix Table 1). Nevertheless, when we tried to intersect our results with the work that was done by Lemischka and Melton we found 3 genes common to our 271 and Lemischka's list, and 11 genes shared with Melton. Moreover, this list of genes does not include the genes mentioned in table 3.1 which belong to cluster H3 or K3 and are known to be involved in self-renewal. This fact is consistent with our claim that the 3-way intersection is not above the level of chance, and casts doubt on the existence of a shared core of genes that control self-renewal in all stem cells. We believe that there are no ”stemness” genes. On the contrary, different stem cell types may use different gene networks to achieve self renewal or pluripotency.

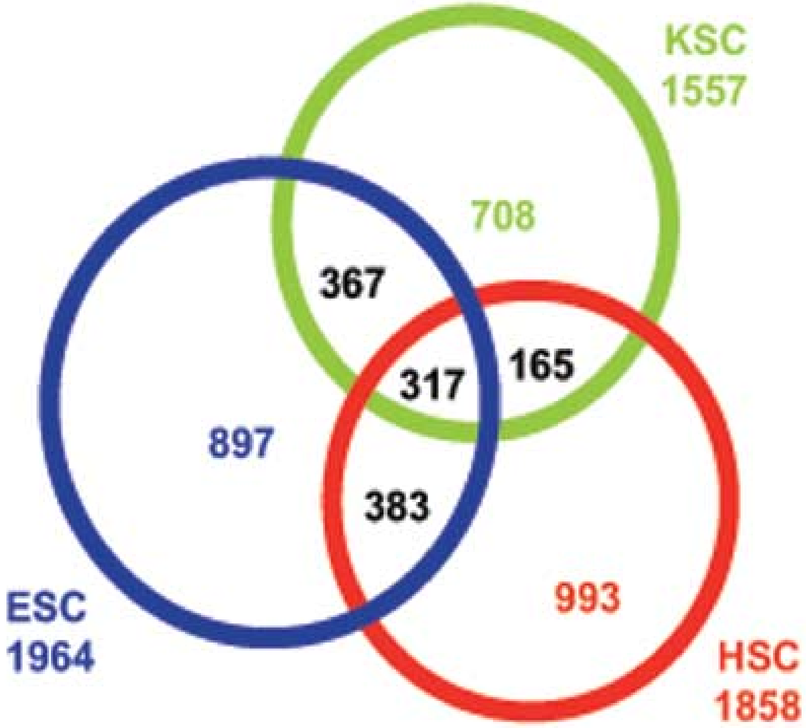


Figure 3.10 Venn diagrams of candidate stemness genes. Intersection of genes enriched in all three Stem Cells (317) as determined by fold change analysis. We selected genes that showed enrichment in KSC (and HSC) by at least 2-fold change of expression compared with their terminally differentiated counterparts KDC (or HDC). ESC enriched genes were selected (2-fold change) over KDC and HDC separately, and then by intersecting Unigene accession numbers.

3.6 Classification Analysis

We further classified the thousands of genes expressed in ESC (genes in Fig. 3.4 and 3.5) into 14 functional categories defined by the Gene Ontology "Biological Process" (http://www.geneontology.org/). We found that the global reduction of expressed genes upon commitment to differentiation (Fig. 3.3) was accompanied by a reduction in the number of transcription factors and a dramatic increase in receptors and cell-cell signaling (Fig. 3.11). We also searched for genes involved in remodeling the chromatin structure because it is very likely that "just in case" strategy is made possible by maintaining an open chromatin structure at the stem cell stage and epigenetic modification upon differentiation [23, 24]. The analysis showed a complete change of a set of genes involved in remodeling the chromatin structure (Fig. 3.12), such as an enrichment in clusters 1-3 (H or K) of helicases of the SWI/SNF family that promote DNA unwinding and enhance transcription. In contrast, cluster 4-5 represent the differentiation state show enrichment in chromatin modifiers that suppress transcription (See full list appendix table 2). The probability (FDR controlled) of observing this enrichment by chance as estimated using hyper geometrical distribution [25] is extremely low.

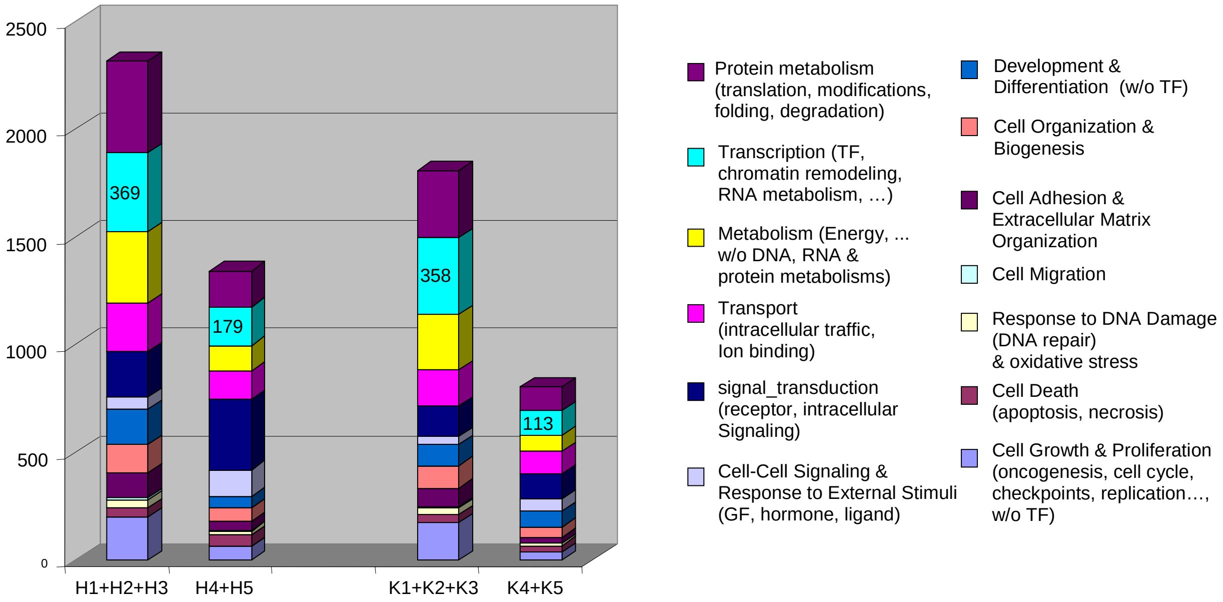


Figure 3.11 Distribution of genes by functional categories (total number w/o ESTs/unknown). The genes from the clusters in Fig. 3.4 and 3.5 were classified by functional categories. Genes were categorized into 14 categories by the Gene Ontology “Biological Process” (http://www.geneontology.org/). The following classification shows reduction in the number of transcription factors while a dramatic increase in receptors and cell-cell signaling.

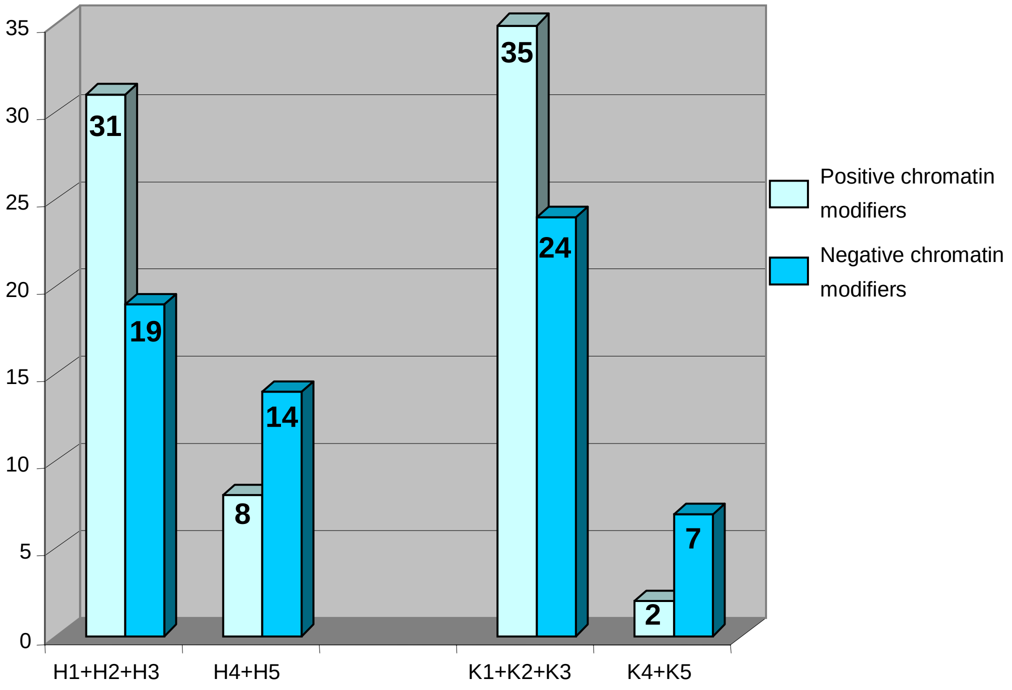


Figure 3.12 Distribution of positive & negative chromatin modifiers. The number of genes from the clusters in Fig. 3.4 and 3.5, responsible for chromatin structure modification, and classified as enhancers and suppressors [23, 24] is shown. P-values of SWI/SNF and HMG genes that fall specifically into clusters 1+2+3 were found to be highly significant in a hyper geometric distribution test (FDR controlled). (SWI/SNF genes: p = 1.0x10-7 in K1+K2+K3, p = 2.8x10-5 in H1+H2+H3; HMG genes: p = 4.5x10-4 in K1+K2+K3, p = 9.7x10-3 in H1+H2+H3).

Reference List

1. Hu, M., et al., *Multilineage gene expression precedes commitment in the hemopoietic system.* Genes Dev, 1997. **11**(6): p. 774-85.
2. Terskikh, A.V., et al., *Gene expression analysis of purified hematopoietic stem cells and committed progenitors.* Blood, 2003. **102**(1): p. 94-101.
3. Akashi, K., et al., *Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis.* Blood, 2003. **101**(2): p. 383-9.
4. Sokal, R.R. and F.J. Rohlf, *Biometry: the principles and practice of statistics in biological research*. 3rd edition ed. 1995, New York: Freeman W. H. and Co. pp. 392.
5. Benjamini, Y.a.H., Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. Roy. Stat. Soc. B., 1995. **57**: p. 289–300.
6. Blatt, M., S. Wiseman, and E. Domany, *Superparamagnetic clustering of data.* Physical Review Letters, 1996. **76**(18): p. 3251-3254.
7. Ivanova, N.B., et al., *A stem cell molecular signature.* Science, 2002. **298**(5593): p. 601-4.
8. Ramalho-Santos, M., et al., *"Stemness": transcriptional profiling of embryonic and adult stem cells.* Science, 2002. **298**(5593): p. 597-600.
9. Nichols, J., et al., *Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4.* Cell, 1998. **95**(3): p. 379-91.
10. Chambers, I., et al., *Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells.* Cell, 2003. **113**(5): p. 643-55.
11. Avilion, A.A., et al., *Multipotent cell lineages in early mouse development depend on SOX2 function.* Genes Dev, 2003. **17**(1): p. 126-40.
12. Sato, N., et al., *Molecular signature of human embryonic stem cells and its comparison with the mouse.* Developmental Biology, 2003. **260**: p. 404-413.
13. Henderson, J.K., et al., *Preimplantation human embryos and embryonic stem cells show comparable expression of stage-specific embryonic antigens.* Stem Cells, 2002. **20**(4): p. 329-37.
14. Pellegrini, G., et al., *p63 identifies keratinocyte stem cells.* Proc Natl Acad Sci U S A, 2001. **98**(6): p. 3156-61.
15. Dowling, J., Q.C. Yu, and E. Fuchs, *Beta4 integrin is required for hemidesmosome formation, cell adhesion and cell survival.* J Cell Biol, 1996. **134**(2): p. 559-72.

16. Tseng, H. and H. Green, *Association of basonuclin with ability of keratinocytes to multiply and with absence of terminal differentiation.* J Cell Biol, 1994. **126**(2): p. 495-506.
17. Park, I.K., et al., *Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells.* Nature, 2003. **423**(6937): p. 302-5.
18. Gering, M., et al., *The SCL gene specifies haemangioblast development from early mesoderm.* Embo J, 1998. **17**(14): p. 4029-45.
19. Orkin, S.H. and L.I. Zon, *Hematopoiesis and stem cells: plasticity versus developmental heterogeneity.* Nat Immunol, 2002. **3**(4): p. 323-8.
20. Su, A.I., et al., *Large-scale analysis of the human and mouse transcriptomes.* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4465-70.
21. Dazard, J.E., et al., *Genome-wide comparison of human keratinocyte and squamous cell carcinoma responses to UVB irradiation: implications for skin and epithelial cancer.* Oncogene, 2003. **22**(19): p. 2993-3006.
22. Wells, J.M., *Genes expressed in the developing endocrine pancreas and their importance for stem cell and diabetes research.* Diabetes Metab Res Rev, 2003. **19**(3): p. 191-201.
23. Razin, A., *CpG methylation, chromatin structure and gene silencing-a three-way connection.* Embo J, 1998. **17**(17): p. 4905-8.
24. Grewal, S.I. and D. Moazed, *Heterochromatin and epigenetic control of gene expression.* Science, 2003. **301**(5634): p. 798-802.
25. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nat Genet, 1999. **22**(3): p. 281-5.

Chapter 4

4 Discussion

In this work we investigated the genetic profile of embryonic and adult stem cells together with their mature progenies. One of the aims of this research was to find common genes to all stem cell tissues in humans. Other groups, led by developmental geneticist Douglas Melton of Harvard University [1] and Ihor Lemischka of Princeton University [2], compared the gene expression of embryonic stem cells, hematopoietic or blood-forming stem cells and neural stem cells in mice. Lemischka's group found 283 genes that were over expressed in all three of their stem cell populations. They interpreted this as indicating that these genes from a part of a genetic characterization of "stemness". Melton's group found 230 genes that were highly expressed in their stem cells. The work of these two groups aimed at identification of stemness signature genes, common to all stem cell types, and could have made a big impact on the use of adult stem cells as part of "personalized genetic therapy" instead of using embryonic stem cells in cell therapy of several diseases. However, the overlap between the two lists of "stemness" genes was very small, which prompted several recently published technical comments and an editorial [3-5] in Science; which exposed the debate around the important question of "stemness": are there genes common to all stem cells? Unfortunately, most information published to date offers more confusion than consensus. The two sets of genes of Lemischka and Melton were almost mutually exclusive, sharing only six genes. A recent study added to the confusion: in a technical comment published online by Science [3], Bing Lim and colleagues at the Genome Institute in Singapore and the Beth Israel Deaconess Medical Center in Boston describe a similar experiment with embryonic stem cells, neural stem cells, and retinal stem cells, also in mice . They found 385 genes that were over expressed in all three cell types.

However, only one of those genes was on both Melton's and Lemischka's lists (see Fig. 4.1).

So what seems to be the problem? Lemischka and Melton proposed several possible reasons for the observed discrepancies. For example that the initial cell population can make a huge difference in what is found in the microarray. "One danger here is that the resolution power of the gene chip technology might be on the verge of outstripping the resolution of the biological assays" for isolation stem cells, Lemischka said. Any genes expressed by partially differentiated cells in the analyzed population will cloud the gene array results. Key genes might vary their expression over time, or perhaps the sought-after stemness genes are absent from the commercially available chips that all three teams used. Lemischka and Melton show [1] [2], however, that when just one stem cell tissue was compared between the three [1-3] studies, a significant number of overlapping genes (with low probability) could be found (Fig. 4.1B, 4.1C). However, when they tried to combine just two types of stem cells, the number of overlapping genes between the three studies was not significant (Fig. 4.1D). Our results show that commitment to a target cell type upon differentiation is accompanied by downregulation of the "inappropriate" genes, i.e. most needed by various tissues but not by the target cell type, and upregulation to dominance of the genes related to the committed target cell type. At the ESC stage pluripotency is maintained by keeping open a large repertoire of gene transcripts, even though they may not be related to maintaining the state of ESC, in anticipation to all options of cell differentiation. At the ASC stage the option for trans or cross differentiation is maintained again by keeping open a repertoire of genes that may not be needed when the ASC is terminally differentiated to a particular cell fate. Therefore, we believe that the cores of "stemeness" genes that were found in each research [1-3], reflect simply the intersection of genes corresponding to the particular all fates that were studied in each case.

In summary, speculations made in independent studies about the identity of “stemness” genes do not hold up when the studies are compared. We believe that the methods used in our study, which included the first use of advanced clustering in this field and which extend far beyond the standard “fold-change and intersect” methods used so far, are a better approach for studying the stemness question.

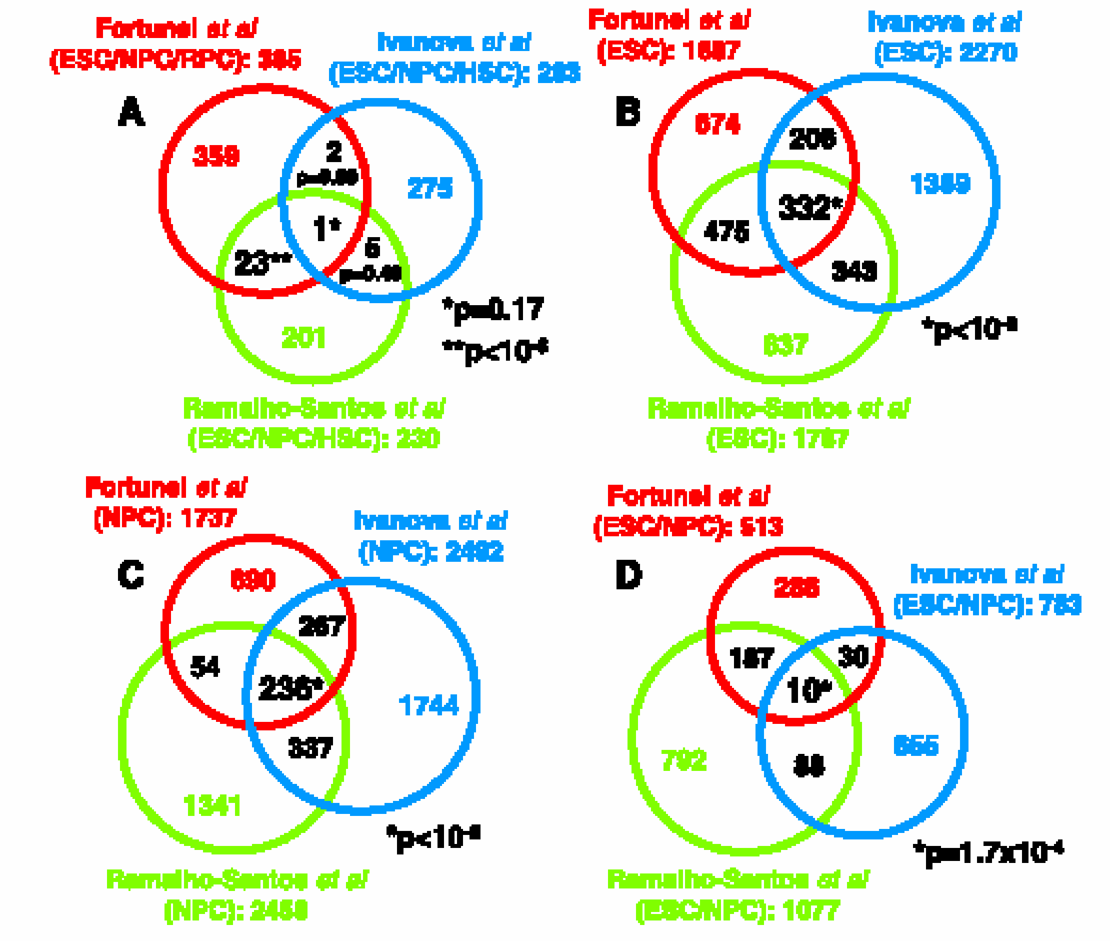


Figure 4.1 Venn diagrams showing overlap of “stemness” genes and stem cell –enriched genes among studies by Ramalho-Santos et al. [1], Ivanova et al. [2], and Fortunel et al . [3] A “Stemness” genes found by the three groups overlap by only one gene. (P =0.17). **B** ESC (Embryonic Stem Cell) -enriched genes identified by each study overlap by 332 genes; the probability that such overlap occurs by chance is extremely low (P $<10^{-8}$). **C** NPC (Neural Progenitor/stem Cells) -enriched genes overlapping by 236 genes between the three groups (P $<10^{-6}$). **D** Overlap of “stemness” genes —two types of stem cell (ESC/NPC)-enriched genes —is limited to 10 genes. The probability of this number of genes overlapping by chance is greatly increased. P > 10^{-4} is not significant because there are more than 10^{4} genes studied.

One of the striking results of our work was that in order to maintain pluripotency, stem cells turn on thousands of genes which represent differentiation pathways into many possible target tissues. Most of these genes are down-regulated upon commitment to a particular cell fate, while genes specific to the target tissue are upregulated. This strategy implies a design principle of stem cells for achieving pluripotency: expressing many genes and then selecting only a few for continued expression as they differentiate, while all other genes will be shut off. This model can help us predict which tissues a specific adult stem cell, e.g. blood stem cell, can differentiate into besides a mature blood cell (Fig. 4.2). This model can help us find the answers to questions like: Are adult SC plastic? Is plasticity selective? A similar model was previously proposed in the case of hematopoietic stem cell differentiation on the basis of expression of erythroid or granulocyte markers in the progenitor cell prior to commitment [6] and recent work extended this also to the analysis of genes in the hematopoietic system [7, 8]. Our study demonstrates the generality of this model and extends it to human ESC and ASC at the level of global gene expression. It is likely that the genes expressed in ESC may also help in choosing the adequate cues that target ESC towards a desired differentiation pathway.

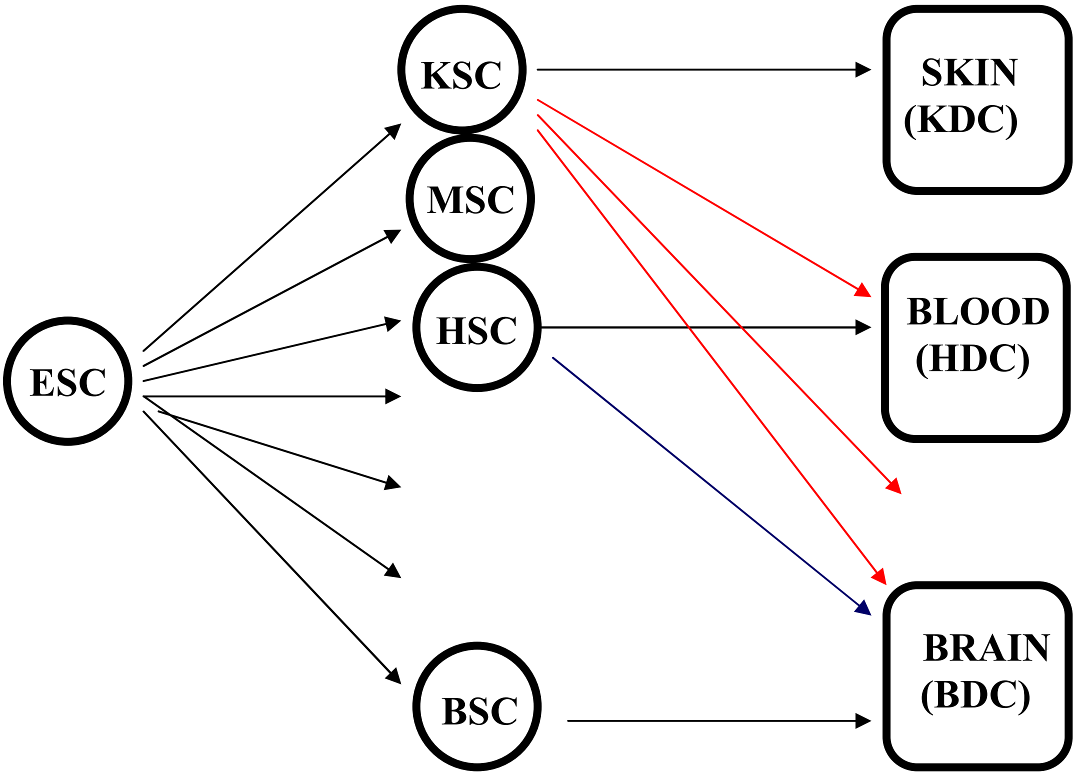


Figure 4.2 Trans-differentiation in Adult Stem Cell. ESC (Embryonic Stem Cells) can differentiate into all ADC (adult stem cell) types (indicated by black arrows). ADC typically generates the cell types of the tissue in which they reside (again, indicated by black arrows). In addition the model allows for trans-differentiation into other tissues (indicated by red and blue arrows) which can be predicted using the genes ADC expresses.

4.1 Future Goals

Functional characterization of candidate central genes for "stemness"

Candidate genes will be cloned and over-expressed in stem cells and the phenotype of the cells will be analyzed and compared with that of untreated stem cells. This will be done using either regulated gene expression or gene "knock-down" by RNA Interference (RNAi). Examples for such genes will be taken from Cluster 3, which probably controls the self-renewal properties (e.g. Nanog).

Defining on/off switching map for a particular pathway of differentiation

Identification of genes that control differentiation is of central importance for various fields in regenerative medicine including gene therapy and tissue engineering [9]. This includes understanding ligand-receptor interaction and the intracellular components of the signaling system, as well as identifying the genes that are activated or inactivated during differentiation of specific cell types [10]. We plan to focus mainly on changes in the repertoire of receptors during differentiation in order to study the above listed questions, because the expression of receptors was found to change significantly between stem to mature cells.

Cancer stem cells

Studies in leukemia demonstrated that only a rare subset of leukemic cells, called "cancer stem cells", possesses the ability to initiate tumor growth [9] [11]. A recent publication on breast cancer conclusively demonstrated that also in this solid tumor only a small subset, of breast cancer stem cells, is capable of initiating and propagating the tumor [12] (See figure 4.3, 4.4). This subset of cancer cells is different from the majority of the tumor both in functionality and by cell surface markers they express. Furthermore, these surface markers are similar to those of the normal stem cell [13]. Major questions in that field are: Are cancer stem cells the target for transforming mutation? Are they also the right target for cancer therapy? What is common for the self-renewal mechanism of ESC and cancer stem cells? Why and how do cancer stem cells disable their differentiation mechanisms and become immortal? This study may change our view on the target for cancer therapy and may open ways for tissue damage repair in "personalized medicine".

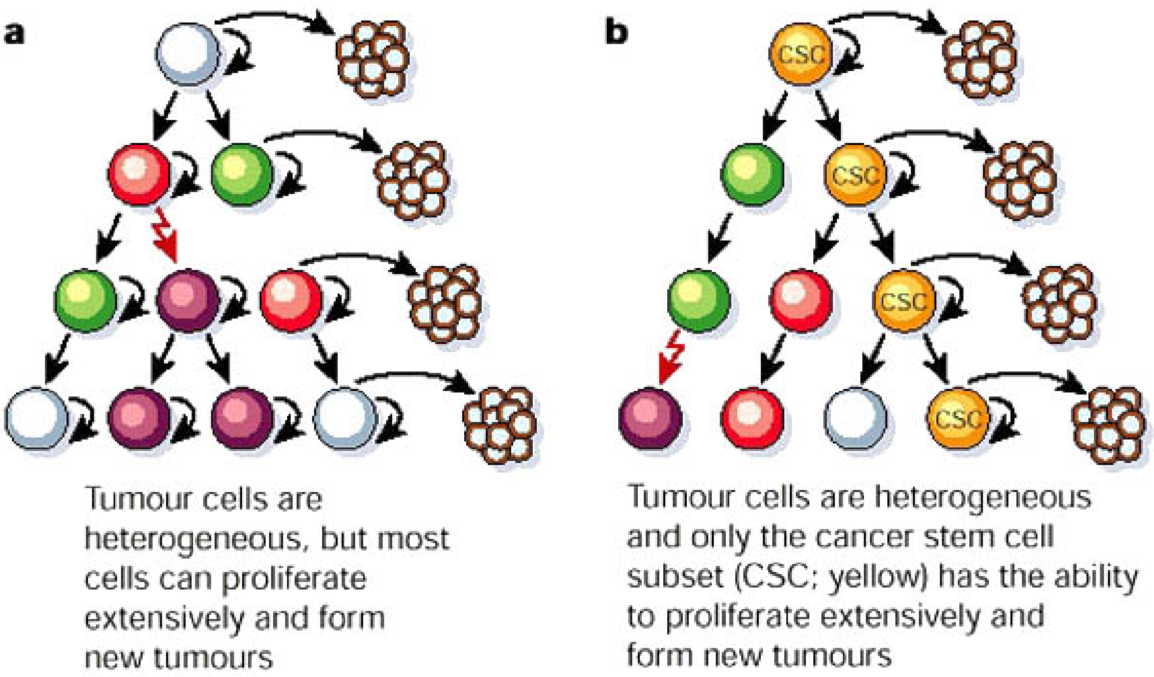

a
Tumour cells are heterogeneous, but most cells can proliferate extensively and form new tumours
b
CSC
CSC
CSC
CSC
Tumour cells are heterogeneous and only the cancer stem cell subset (CSC; yellow) has the ability to proliferate extensively and form new tumours

Figure 4.3 Two general models of heterogeneity in solid cancer cells a, Cancer cells of many different phenotypes have the potential to proliferate extensively, but any one cell would have a low probability of exhibiting this potential in an assay of clonogenicity or tumorigenicity. **b,** Most cancer cells have only limited proliferative potential, but a subset of cancer cells consistently proliferate extensively in clonogenic assays and can form new tumors on transplantation. The model shown in b predicts that a distinct subset of cells is enriched for the ability to form new tumors, whereas most cells are depleted of this ability. Existing therapeutic approaches have been based largely on the model shown in a, but the failure of these therapies to cure most solid cancers suggests that the model shown in b may be more accurate.

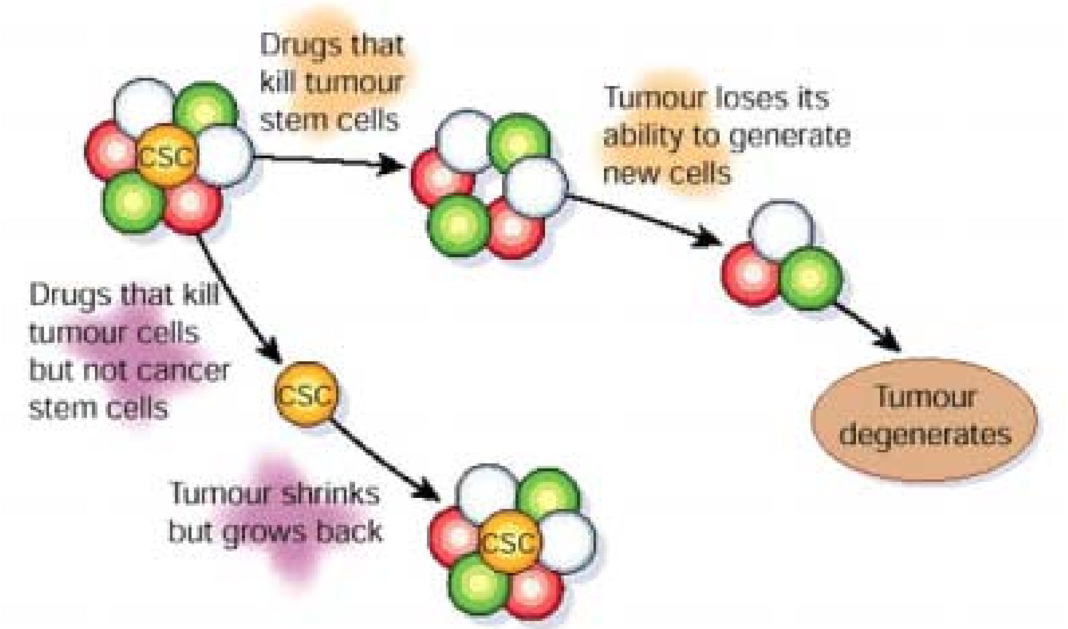

Drugs that kill tumour stem cells
CSC
Tumour loses its ability to generate new cells
Drugs that kill tumour cells but not cancer stem cells
CSC
Tumour degenerates
Tumour shrinks but grows back
CSC

Figure 4.4 Conventional therapies may shrink tumors by killing mainly cells with limited proliferation potential. If the putative cancer stem cells are less sensitive to these therapies, then they will remain viable after therapy and re-establish the tumour. By contrast, if therapies can be targeted against cancer stem cells, they might more effectively kill the cancer stem cells, rendering the tumours unable to maintain themselves or grow. Thus, even if cancer stem cell-directed therapies do not shrink tumours initially, they may eventually lead to cures.

Reference List

1. Ramalho-Santos, M., et al., *"Stemness": transcriptional profiling of embryonic and adult stem cells.* Science, 2002. **298**(5593): p. 597-600.
2. Ivanova, N.B., et al., *A stem cell molecular signature.* Science, 2002. **298**(5593): p. 601-4.
3. Fortunel, N.O., et al., *Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature".* Science, 2003. **302**(5644): p. 393; author reply 393.
4. Evsikov, A.V. and D. Solter, *Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature".* Science, 2003. **302**(5644): p. 393; author reply 393.
5. Vogel, G., *Stem cells. 'Stemness' genes still elusive.* Science, 2003. **302**(5644): p. 371.
6. Hu, M., et al., *Multilineage gene expression precedes commitment in the hemopoietic system.* Genes Dev, 1997. **11**(6): p. 774-85.
7. Terskikh, A.V., et al., *Gene expression analysis of purified hematopoietic stem cells and committed progenitors.* Blood, 2003. **102**(1): p. 94-101.
8. Akashi, K., et al., *Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis.* Blood, 2003. **101**(2): p. 383-9.
9. Turgeman, G., et al., *Engineered human mesenchymal stem cells: a novel platform for skeletal cell mediated gene therapy.* J Gene Med, 2001. **3**(3): p. 240-51.
10. Odorico, J.S., D.S. Kaufman, and J.A. Thomson, *Multilineage differentiation from human embryonic stem cell lines.* Stem Cells, 2001. **19**(3): p. 193-204.
11. Bhatia, R. and P.B. McGlave, *Autologous stem cell transplantation for the treatment of chronic myelogenous leukemia.* Cancer Treat Res, 1997. **77**: p. 357-74.
12. Dick, J.E., *Breast cancer stem cells revealed.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 3547-9.
13. Al-Hajj, M., et al., *Prospective identification of tumorigenic breast cancer cells.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 3983-8.

5 Summary

1. Human embryonic stem cells (ESC) are undifferentiated and are endowed with the capacities of self renewal and pluripotential differentiation. Adult stem cells (ASC) renew their own tissue, but whether they can transdifferentiate to other tissues is still debated. To understand the genetic program that underlies the functioning of stem cells, we set out to determine whether there exists a common core of so-called "stemness" genes, shared by all stem cells (SC), which accounts for both self-renewal and pluripotency. To this end we compared the transcriptomes of ESC with ASC of human hematopoietic (HSC) and keratinocytic (KSC) origins, along with their mature progenies.
2. Using advanced clustering, we divided the genes according to the transcriptomes or genetic profiles of each differentiation pathway (Hemapotietic and Keratinocyte). By comparing the genes which their biological functions have been known from the literature, we suggested that there are no shared “stemness” genes. Rather, there are two different groups of genes, one related to self renewal and the other to pluripotency. The genes related to self renewal are specific to each SC type and different from those of ESC.
3. Another group of genes common to ESC and ASC appear to be related to pluripotency and plasticity of adult stem cells. In order to maintain pluripotency, stem cells turn on thousands of genes which represent differentiation pathways into many possible target tissues. Most of these genes are down-regulated upon commitment to a particular cell fate, while genes specific to the target tissue are upregulated.
4. This strategy implies a design principle model of stem cells for achieving pluripotency; expressing many genes and then selecting only a few for continued expression as they differentiate.

5. We think that these results are of great interest to scientists from many fields and will help to shed light on important controversies in stem cells research. This will fertilize new ideas for future research based on gene programming in stem cells.

6 Appendix

6.1 Appendix Table 1

List of 317 candidate stemness genes common to ESC and ASC Genes were categorized into 14 categories by the Gene Ontology "Biological Process" (http://www.geneontology.org/). GO term refers to the name(s) of each category, and GO ID to its (their) identification number(s). In addition to fold change analysis two statistical tests, t-test and Wilcoxon Rank Sum test were used to select genes that expressed differentially in stem cells. Intersection of this list with the 317 genes yielded 271 genes common to all methods (indicated by stars (*)).

16049 / 8283 - Cell Growth & Proliferation (oncogenesis, cell cycle, checkpoints, replication…,w/o TF)

1	NM_020993.1	BCL7A	B-cell CLL/lymphoma 7A	*
2	NM_004642.1	CDK2AP1	CDK2-associated protein 1	*
3	NM_001274.1	CHEK1	CHK1 checkpoint homolog (S. pombe)	*
4	AF234161.1	CIZ1	Cip1-interacting zinc finger protein	*
5	NM_000075.1	CDK4	cyclin-dependent kinase 4	*
6	AF321125.1	CDT1	DNA replication factor	*
7	AI924630	MAGED2	cDNA highly similar to Human hepatocellular carcinoma associated protein (JCL-1)	
8	NM_014708.1	KNTC1	kinetochore associated 1	*
9	NM_016073.1	HDGFRP3	likely ortholog of mouse hepatoma-derived growth factor, related protein 3	*
10	NM_022149.1	MAGEF1	MAGEF1 protein	*
11	AF217963.1	MAGED1	melanoma antigen, family D, 1	*
12	AF126181.1	MAGED2	melanoma antigen, family D, 2	*
13	NM_004526.1	MCM2	MCM2 minichromosome maintenance deficient 2, mitotin (S. cerevisiae)	*
14	NM_002388.2	MCM3	MCM3 minichromosome maintenance deficient 3 (S. cerevisiae)	*
15	X74794.1	MCM4	MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)	*
16	AA807529	MCM5	MCM5 minichromosome maintenance deficient 5, cell division cycle 46 (S. cerevisiae)	*
17	NM_005915.2	MCM6	MCM6 minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe) (S. cerevisiae)	*
18	D55716.1	MCM7	MCM7 minichromosome maintenance deficient 7 (S. cerevisiae)	*
19	U04045.1	MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	*
20	D89646.1	MSH6	mutS homolog 6 (E. coli)	*
21	NM_014303.1	PES1	pescadillo homolog 1, containing BRCT domain (zebrafish)	*
22	NM_016937.1	POLA	polymerase (DNA directed), alpha	*
23	NM_006230.1	POLD2	polymerase (DNA directed), delta 2, regulatory subunit 50kDa	*
24	NM_002692.1	POLE2	polymerase (DNA directed), epsilon 2 (p59 subunit)	*
25	NM_000946.1	PRIM1	primase, polypeptide 1, 49kDa	*
26	AL560017	PHB	prohibitin	*
27	NM_006443.1	RCL	putative c-Myc-responsive	*
28	NM_006397.1	RNASEH2A	ribonuclease H2, large subunit	*
29	M93651	SET	SET translocation (myeloid leukemia-associated)	*

8219 - Cell Death (apoptosis, necrosis, autophagy)

1	NM_005087.1	FXR1	fragile X mental retardation, autosomal homolog 1	*
2	AI336206	PAWR	PRKC, apoptosis, WT1, regulator	*
3	NM_016629.1	TNFRSF21	tumor necrosis factor receptor superfamily, member 21	*

6974 / 6979Response to DNA Damage (DNA repair, …) & Oxidative Stress

1	M32721.1	ADPRT	ADP-ribosyltransferase (NAD+; poly (ADP-ribose) polymerase)	
2	M80261.1	APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	*
3	NM_000465.1	BARD1	BRCA1 associated RING domain 1	*
4	NM_000178.1	GSS	glutathione synthetase	*
5	NM_005590.1	MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)	*
6	NM_002452.1	NUDT1	nudix (nucleoside diphosphate linked moiety X)-type motif 1	*
7	NM_006406.1	PRDX4	peroxiredoxin 4	*
8	NM_005732.1	RAD50	RAD50 homolog (S. cerevisiae)	
9	NM_005410.1	SEPP1	selenoprotein P, plasma, 1	*
10	NM_003362.1	UNG	uracil-DNA glycosylase	*

7155 / 30198 - Cell Adhesion & Extracellular Matrix Organization

1	NM_000484.1	APP	amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)	*
2	NM_004357.1	CD151	CD151 antigen	
3	AW052179	COL4A5	collagen, type IV, alpha 5 (Alport syndrome)	
4	AI983428	COL5A1	collagen, type V, alpha 1	*
5	NM_014288.1	ITGB3BP	integrin beta 3 binding protein (beta3-endonexin)	*
6	NM_002293.2	LAMC1	laminin, gamma 1 (formerly LAMB2)	*
7	NM_003628.2	PKP4	plakophilin 4	*
8	NM_000297.1	PKD2	polycystic kidney disease 2 (autosomal dominant)	*
9	NM_005505.1	SCARB1	scavenger receptor class B, member 1	*

16043 - Cell Organization and Biogenesis (cytoskeleton, …)

1	AL533838	OK/SW-cl.56	beta 5-tubulin	*
2	BC004912.1	BPAG1	bullous pemphigoid antigen 1, 230/240kDa	*
3	L07515.1	CBX5	chromobox homolog 5 (HP1 alpha homolog, Drosophila)	
4	NM_004395.1	DBN1	drebrin 1	*
5	NM_006824.1	EBNA1BP2	EBNA1 binding protein 2	*
6	NM_005886.1	KATNB1	katanin p80 (WD40-containing) subunit B 1	*
7	M94363	LMNB2	lamin B2	*
8	AK026977.1	MYH10	myosin, heavy polypeptide 10, non-muscle	*
9	NM_014502.1	NMP200	nuclear matrix protein NMP200 related to splicing factor PRP19	*
10	NM_006985.1	NPIP	nuclear pore complex interacting protein	*
11	NM_006993.1	NPM3	nucleophosmin/nucleoplasmin, 3	*
12	AL162068.1	NAP1L1	nucleosome assembly protein 1-like 1	*
13	NM_006444.1	SMC2L1	SMC2 structural maintenance of chromosomes 2-like 1 (yeast)	*
14	AL136877.1	SMC4L1	SMC4 structural maintenance of chromosomes 4-like 1 (yeast)	
15	BE968833	SPTBN1	spectrin, beta, non-erythrocytic 1	*
16	NM_005563.2	STMN1	stathmin 1/oncoprotein 18	*
17	AC004472	STOML2	stomatin (EPB72)-like 2	*
18	NM_003289.1	TPM2	tropomyosin 2 (beta)	*
19	NM_006082.1	K-ALPHA-1	tubulin, alpha, ubiquitous	*
20	NM_005775.1	SCAM-1	vinexin beta (SH3-containing adaptor molecule-1)	*

7275 / 30154 - Development & Differentiation (w/o TF)

1	Y15521	CRIP2	cysteine-rich protein 2	*
2	NM_004939.1	DDX1	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 1	*
3	NM_001449.1	FHL1	four and a half LIM domains 1	*
4	BC000915.1	PDLIM1	PDZ and LIM domain 1 (elfin)	
5	NM_002482.1	NASP	nuclear autoantigenic sperm protein (histone-binding)	*
6	NM_006623.1	PHGDH	phosphoglycerate dehydrogenase	*
7	NM_002573.1	PAFAH1B3	platelet-activating factor acetylhydrolase, isoform Ib, gamma subunit 29kDa	*
8	NM_003877.1	SOCS2	suppressor of cytokine signaling 2	*

7267 / 9605 - Cell-Cell Signaling & Response to External Stimulus (GF, hormone, ligand,)

1	BC000055.1	FSTL1	follistatin-like 1	*
2	NM_001553.1	IGFBP7	insulin-like growth factor binding protein 7	
3	NM_016205.1	PDGFC	platelet derived growth factor C	*
4	NM_000062.1	SERPING1	serine (or cysteine) proteinase inhibitor, clade G (C1 inhibitor), member 1, (angioedema, hereditary)	*

7165 - Signal Transduction (receptor, intracellular signaling, …)

1	NM_004444.1	EPHB4	EphB4	
2	M37712.1	GPR125	G protein-coupled receptor 125	*
3	NM_006055.1	LANCL1	LanC lantibiotic synthetase component C-like 1 (bacterial)	*
4	BE879873	PGRMC2	progesterone receptor membrane component 2	
5	NM_002821.1	PTK7	PTK7 protein tyrosine kinase 7	*
6	NM_002882.2	RANBP1	RAN binding protein 1	*
7	AF015043.1	SH3BP4	SH3-domain binding protein 4	*
8	AW131863	SH3GLB2	SH3-domain GRB2-like endophilin B2	
9	NM_022748.1	TEM6	tumor endothelial marker 6	*
10	NM_003931.1	WASF1	WAS protein family, member 1	*

6810 - Transport (intracellular traffic, ion binding, …)

1	AI002002	ABCE1	ATP-binding cassette, sub-family E (OABP), member 1	*
2	AF005422.1	CLNS1A	chloride channel, nucleotide-sensitive, 1A	*
3	BE256479	HSPD1	heat shock 60kDa protein 1 (chaperonin)	*
4	AI144007	HNRPA1	heterogeneous nuclear ribonucleoprotein A1	
5	NM_024658.1	IPO4	importin 4	*
6	NM_018085.1	IPO9	importin 9	*
7	NM_002271.1	KPNB3	karyopherin (importin) beta 3	*
8	NM_018230.1	NUP133	nucleoporin 133kDa	*
9	NM_024647.1	NUP43	nucleoporin Nup43	*
10	NM_006227.1	PLTP	phospholipid transfer protein	*
11	NM_004955.1	SLC29A1	solute carrier family 29 (nucleoside transporters), member 1	*
12	NM_014765.1	TOMM20	translocase of outer mitochondrial membrane 20 (yeast) homolog	*
13	N36842	UPF3A	UPF3 regulator of nonsense transcripts homolog A (yeast)	*

8152 - Metabolism (Energy, ...w/o DNA, RNA & protein metabolisms)

1	AW000964	HIBCH	3-hydroxyisobutyryl-Coenzyme A hydrolase	*
2	D89976.1	ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	*
3	BE855983	ACACA	acetyl-Coenzyme A carboxylase alpha	
4	AF067854.1	ADSL	adenylosuccinate lyase	
5	M30471.1	ADH5	alcohol dehydrogenase 5 (class III), chi polypeptide	*
6	AF130089.1	ALDH6A1	aldehyde dehydrogenase 6 family, member A1	*
7	BC002515.1	ALDH7A1	aldehyde dehydrogenase 7 family, member A1	*
8	AB009598	B3GAT3	beta-1,3-glucuronyltransferase 3 (glucuronosyltransferase I)	*
9	NM_004341.1	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	*
10	L35594.1	ENPP2	ectonucleotide pyrophosphatase/phosphodiesterase 2 (autotaxin)	
11	NM_001428.1	ENO1	enolase 1, (alpha)	*
12	BE540552	FADS1	fatty acid desaturase 1	*
13	NM_004265.1	FADS2	fatty acid desaturase 2	*
14	NM_001512.1	GSTA4	glutathione S-transferase A4	
15	NM_000156.3	GAMT	guanidinoacetate N-methyltransferase	*
16	NM_016576.1	GMPR2	guanosine monophosphate reductase 2	*

17	NM_000194.1	HPRT1	hypoxanthine phosphoribosyltransferase 1 (Lesch-Nyhan syndrome)	*
18	NM_000884.1	IMPDH2	IMP (inosine monophosphate) dehydrogenase 2	*
19	NM_002300.1	LDHB	lactate dehydrogenase B	*
20	NM_002402.1	MEST	mesoderm specific transcript homolog (mouse)	*
21	BC001686.1	MAT2A	methionine adenosyltransferase II, alpha	*
22	NM_005956.2	MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent), methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase	*
23	NM_000269.1	NME1	non-metastatic cells 1, protein (NM23A) expressed in	*
24	NM_002512.1	NME2	non-metastatic cells 2, protein (NM23B) expressed in	*
25	NM_013330.2	NME7	non-metastatic cells 7, protein expressed in (nucleoside-diphosphate kinase)	*
26	L14599.1	NONO	non-POU domain containing, octamer-binding	*
27	NM_006117.1	PECI	peroxisomal D3,D2-enoyl-CoA isomerase	*
28	NM_007169.1	PEMT	phosphatidylethanolamine N-methyltransferase	*
29	U24183.1	PFKM	phosphofructokinase, muscle	*
30	U00238.1	PPAT	phosphoribosyl pyrophosphate amidotransferase	*
31	AA902652	PAICS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazole succinocarboxamide synthetase	*
32	Y09703.1	PNN	pinin, desmosome associated protein	
33	NM_002860.1	PYCS	pyrroline-5-carboxylate synthetase (glutamate gamma-semialdehyde synthetase)	*
34	NM_000687.1	AHCY	S-adenosylhomocysteine hydrolase	*
35	NM_014285.1	RRP4	homolog of Yeast RRP4 (ribosomal RNA processing 4), 3'-5'-exoribonuclease	*
36	NM_003089.1	SNRP70	small nuclear ribonucleoprotein 70kDa polypeptide (RNP antigen)	*
37	BC001721.1	SNRPD1	small nuclear ribonucleoprotein D1 polypeptide 16kDa	*
38	NM_003094.1	SNRPE	small nuclear ribonucleoprotein polypeptide E	*
39	NM_003132.1	SRM	spermidine synthase	
40	NM_001071.1	TYMS	thymidylate synthetase	*
41	NM_006297.1	XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1	*

6350- Transcription (TF, chromatin remodeling, RNA metabolism, …)

1	BC006259.1	CYLN2	cytoplasmic linker 2	*
2	AA485440	DBP	D site of albumin promoter (albumin D-box) binding protein	
3	U59151.1	DKC1	dyskeratosis congenita 1, dyskerin	*
4	NM_000137.1	FAH	fumarylacetoacetate hydrolase (fumarylacetoacetase)	*
5	NM_015487.1	GEMIN4	gem (nuclear organelle) associated protein 4	*
6	NM_001517.1	GTF2H4	general transcription factor IIH, polypeptide 4, 52kDa	*
7	X86401.1	GATM	glycine amidinotransferase (L-arginine:glycine amidinotransferase)	
8	AF274949.1	HMGN3	high mobility group nucleosomal binding domain 3	
9	BE311760	HMGB1	high-mobility group box 1	
10	NM_006709.1	BAT8	HLA-B associated transcript 8	*
11	NM_006559.1	KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1	*
12	AF196468.1	LSM2	LSM2 homolog, U6 small nuclear RNA associated (S. cerevisiae)	*
13	NM_000381.1	MID1	midline 1 (Opitz/BBB syndrome)	
14	U35139.1	NDN	necdin homolog (mouse)	*
15	AF063020.1	PSIP2	PC4 and SFRS1 interacting protein 2	
16	AF268615.1	POU5F1	POU 5 domain protein [Homo sapiens], mRNA sequence	
17	NM_006445.1	PRPF8	PRP8 pre-mRNA processing factor 8 homolog (yeast)	*
18	NM_013235.1	RNASE3L	putative ribonuclease III	*
19	NM_003707.1	RUVBL1	RuvB-like 1 (E. coli)	*

20	NM_006666.1	RUVBL2	RuvB-like 2 (E. coli)	*
21	AF077048.1	SSBP2	single-stranded DNA binding protein 2	*
22	NM_004596.1	SNRPA	small nuclear ribonucleoprotein polypeptide A	*
23	NM_003107.1	SOX4	SRY (sex determining region Y)-box 4	*
24	BE795648	SSRP1	structure specific recognition protein 1	*
25	NM_003069.1	SMARCA1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 1	*
26	NM_003079.1	SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	*
27	AL050318	TGIF2	TGFB-induced factor 2 (TALE family homeobox)	*
28	NM_003195.1	TCEA2	transcription elongation factor A (SII), 2	*
29	M31222.1	TCF3	transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	*
30	K03199.1	TP53	tumor protein p53 (Li-Fraumeni syndrome)	*
31	AB002330.1	SR140	U2-associated SR140 protein	*
32	AF083389.1	WHSC1	Wolf-Hirschhorn syndrome candidate 1	*
33	NM_007145.1	ZNF146	zinc finger protein 146	*

19538 - Protein Metabolism (translation, modifications, folding, degradation, …)

1	AK001980.1	ADPRTL2	ADP-ribosyltransferase (NAD+; poly(ADP-ribose) polymerase)-like 2	*
2	W87689	G2AN	alpha glucosidase II alpha subunit	*
3	NM_000666.1	ACY1	aminoacylase 1	*
4	NM_012100.1	DNPEP	aspartyl aminopeptidase	*
5	NM_001349.1	DARS	aspartyl-tRNA synthetase	*
6	NM_000386.1	BLMH	bleomycin hydrolase	*
7	AU145941	CDC14B	CDC14 cell division cycle 14 homolog B (S. cerevisiae)	*
8	NM_014826.1	CDC42BPA	CDC42 binding protein kinase alpha (DMPK-like)	*
9	AL545982	CCT2	chaperonin containing TCP1, subunit 2 (beta)	*
10	AL078459	DDAH1	dimethylarginine dimethylaminohydrolase 1	
11	NM_002824.1	FKBP4	FK506 binding protein 4, 59kDa	*
12	NM_004667.2	HERC2	hect domain and RLD 2	*
13	NM_001536.1	HRMT1L2	HMT1 hnRNP methyltransferase-like 2 (S. cerevisiae)	*
14	NM_013417.1	IARS	isoleucine-tRNA synthetase	*
15	NM_017840.1	MRPL16	mitochondrial ribosomal protein L16	
16	BC003375.1	MRPL3	mitochondrial ribosomal protein L3	*
17	NM_015956.1	MRPL4	mitochondrial ribosomal protein L4	*
18	AB049636.1	MRPL9	mitochondrial ribosomal protein L9	*
19	NM_016034.1	MRPS2	mitochondrial ribosomal protein S2	*
20	D87453.1	MRPS27	mitochondrial ribosomal protein S27	*
21	NM_016071.1	MRPS33	mitochondrial ribosomal protein S33	*
22	NM_002453.1	MTIF2	mitochondrial translational initiation factor 2	*
23	NM_015909.1	NAG	neuroblastoma-amplified protein	
24	NM_021079.1	NMT1	N-myristoyltransferase 1	
25	NM_004279.1	PMPCB	peptidase (mitochondrial processing) beta	
26	NM_004564.1	PET112L	PET112-like (yeast)	*
27	AD000092	FARSL	phenylalanine-tRNA synthetase-like	*
28	NM_021154.1	PSA	phosphoserine aminotransferase	
29	NM_006451.1	PAIP1	polyadenylate binding protein-interacting protein 1	*
30	NM_014241.1	PTPLA	protein tyrosine phosphatase-like (proline instead of catalytic arginine), member a	*
31	AF009205.1	ARHGEF10	Rho guanine nucleotide exchange factor (GEF) 10	*
32	AL541302	SERPINE2	serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2	*

33	NM_003321.1	TUFM	Tu translation elongation factor, mitochondrial	*
34	NM_003940.1	USP13	ubiquitin specific protease 13 (isopeptidase T-3)	*
35	BC003556.1	USP14	ubiquitin specific protease 14 (tRNA-guanine transglycosylase)	*
36	NM_004481.2	GALNT2	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 2 (GalNAc-T2)	*
37	NM_006295.1	VARS2	valyl-tRNA synthetase 2	*

Other / ESTs / Unknown

1	NM_016608.1	ALEX1	ALEX1 protein	*
2	NM_017797.1	BTBD2	BTB (POZ) domain containing 2	*
3	AI422099	CHD1L	chromodomain helicase DNA binding protein 1-like	
4	NM_006589.1	C1orf2	chromosome 1 open reading frame 2	*
5	NM_016183.1	C1orf33	chromosome 1 open reading frame 33	*
6	NM_004649.1	C21orf33	chromosome 21 open reading frame 33	*
7	NM_018944.1	C21orf45	chromosome 21 open reading frame 45	*
8	NM_004772.1	C5orf13	chromosome 5 open reading frame 13	
9	AW008531	C7orf14	chromosome 7 open reading frame 14	*
10	AW089673	LUC7A	cisplatin resistance-associated overexpressed protein	
11	AU151801	C1QBP	complement component 1, q subcomponent binding protein	*
12	NM_018204.1	CKAP2	cytoskeleton associated protein 2	*
13	AL050022.1	DKFZP564D116	DKFZP564D116 protein	*
14	AL050028.1	DKFZP566C0424	DKFZP566C0424 protein	*
15	AU158148	DKFZP586L0724	DKFZP586L0724 protein	*
16	NM_006014.1	DXS9879E	DNA segment on chromosome X (unique) 9879 expressed sequence	*
17	BF240590	DNAJC9	DnaJ (Hsp40) homolog, subfamily C, member 9	*
18	NM_003720.1	DSCR2	Down syndrome critical region gene 2	*
19	NM_018127.2	ELAC2	elaC homolog 2 (E. coli)	*
20	NM_021178.1	HEI10	enhancer of invasion 10	*
21	AU145746	ESD	esterase D/formylglutathione hydrolase	*
22	BE673445	---	ESTs, Weakly similar to Solute carrier family 11 member 1 (natural resistance-associated macrophage protein 1); Natural resistance-associated macrophage protein; solute carrier family 11 (proton-coupled divalent metal ion transporters), member 1 [Rattus norvegicus] [R.norvegicus]	*
23	AF000416.1	EXTL2	exostoses (multiple)-like 2	*
24	NM_022372.1	GBL	G protein beta subunit-like	
25	AK021980.1	---	Homo sapiens cDNA FLJ11918 fis, clone HEMBB1000272.	
26	BG391282	---	Homo sapiens cDNA FLJ31079 fis, clone HSYRA2001595.	*
27	BG105365	---	Homo sapiens cDNA: FLJ22571 fis, clone HSI02239.	*
28	AW293356	---	Homo sapiens cDNA: FLJ23005 fis, clone LNG00396, highly similar to AF055023 Homo sapiens clone 24723 mRNA sequence.	*
29	BE867771	---	Homo sapiens mRNA; cDNA DKFZp686N1377 (from clone DKFZp686N1377)	
30	BF791738	---	Homo sapiens PRO2751 mRNA, complete cds	*
31	BF967998	---	Homo sapiens, clone IMAGE:5288080, mRNA	*
32	BC003186.1	LOC51659	HSPC037 protein	*
33	BF031714	HYA22	HYA22 protein	*

34	AK001389.1	DKFZP564O043	hypothetical protein DKFZp564O043	*
35	NM_030800.1	DKFZP564O1664	hypothetical protein DKFZp564O1664	*
36	NM_017975.1	FLJ10036	hypothetical protein FLJ10036	*
37	NM_018034.1	FLJ10233	hypothetical protein FLJ10233	*
38	NM_018128.1	FLJ10534	hypothetical protein FLJ10534	*
39	AF274950.1	FLJ10637	hypothetical protein FLJ10637	*
40	AL109978.1	FLJ10737	hypothetical protein FLJ10737	*
41	NM_024662.1	FLJ10774	hypothetical protein FLJ10774	*
42	AL534972	FLJ10849	hypothetical protein FLJ10849	*
43	NM_018268.1	FLJ10904	hypothetical protein FLJ10904	*
44	AA292789	FLJ11029	hypothetical protein FLJ11029	
45	NM_018359.1	FLJ11200	hypothetical protein FLJ11200	*
46	NM_025155.1	FLJ11848	hypothetical protein FLJ11848	*
47	NM_022908.1	FLJ12442	hypothetical protein FLJ12442	*
48	NM_031206.1	FLJ12525	hypothetical protein FLJ12525	
49	NM_017735.1	FLJ20272	hypothetical protein FLJ20272	*
50	NM_017802.1	FLJ20397	hypothetical protein FLJ20397	*
51	NM_019042.1	FLJ20485	hypothetical protein FLJ20485	*
52	NM_017867.1	FLJ20534	hypothetical protein FLJ20534	*
53	NM_022743.1	FLJ21080	hypothetical protein FLJ21080	*
54	NM_024863.1	FLJ21174	hypothetical protein FLJ21174	*
55	NM_024622.1	FLJ21901	hypothetical protein FLJ21901	
56	NM_024678.1	FLJ23441	hypothetical protein FLJ23441	*
57	AI560455	LOC284106	hypothetical protein LOC284106	
58	L19183.1	MAC30	hypothetical protein MAC30	*
59	NM_024113.1	MGC4707	hypothetical protein MGC4707	*
60	U79260.1	MGC5149	hypothetical protein MGC5149	*
61	NM_024096.1	MGC5627	hypothetical protein MGC5627	*
62	NM_018096.1	FLJ10458	hypothetical protein similar to beta-transducin family	*
63	NM_003685.1	KHSRP	KH-type splicing regulatory protein (FUSE binding protein 2)	*
64	D31887.1	KIAA0062	KIAA0062 protein	
65	D42044.1	KIAA0090	KIAA0090 protein	*
66	NM_014669.1	KIAA0095	KIAA0095 gene product	*
67	NM_014641.1	NFBD1	KIAA0170 gene product	*
68	NM_021067.1	KIAA0186	KIAA0186 gene product	*
69	NM_014753.1	KIAA0187	KIAA0187 gene product	*
70	AW205215	KIAA0286	KIAA0286 protein	*
71	NM_014675.1	KIAA0445	KIAA0445 gene product	
72	AB011087.1	KIAA0515	KIAA0515 protein	*
73	AB011154.1	KIAA0582	KIAA0582 protein	*
74	AB011173.1	KIAA0601	KIAA0601 protein	*
75	AI978623	KIAA0657	KIAA0657 protein	*
76	AI493119	KIAA1196	KIAA1196 protein	*

77	NM_025081.1	KIAA1305	KIAA1305 protein	*
78	BC002477.1	KIAA1630	KIAA1630 protein	*
79	NM_003573.1	LTBP4	latent transforming growth factor beta binding protein 4	*
80	NM_016202.1	LOC51157	LDL induced EC protein	*
81	M92439.1	LRPPRC	leucine-rich PPR-motif containing	*
82	NM_014174.1	THY28	likely ortholog of the mouse thymocyte protein Thy28	*
83	NM_018407.1	LAPTM4B	lysosomal associated protein transmembrane 4 beta	*
84	NM_021820.1	MDS024	MDS024 protein	*
85	NM_014878.1	KIAA0020	minor histocompatibility antigen HA-8 (pumilio family)	*
86	NM_022362.1	MMS19L	MMS19-like (MET18 homolog, S. cerevisiae)	*
87	NM_002475.1	MLC1SA	myosin light chain 1 slow a	*
88	BC004944.1	PLINP-1	papillomavirus L2 interacting nuclear protein 1	
89	NM_014051.1	PTD011	PTD011 protein	*
90	NM_016448.1	RAMP	RA-regulated nuclear matrix-associated protein	*
91	NM_002902.1	RCN2	reticulocalbin 2, EF-hand calcium binding domain	*
92	AL049748	RBM9	RNA binding motif protein 9	
93	AI452524	RBMX	RNA binding motif protein, X chromosome	*
94	NM_017512.1	HSRTSBETA	rTS beta protein	*
95	NM_014575.1	SCHIP1	schwannomin interacting protein 1	*
96	AW136988	SSX2IP	synovial sarcoma, X breakpoint 2 interacting protein	*
97	AB020636.1	TIP120A	TBP-interacting protein	*
98	NM_021992.1	TMSNB	thymosin, beta, identified in neuroblastoma cells	*
99	BC001648.1	WDR18	WD repeat domain 18	*
100	NM_018181.1	FLJ10697	zinc finger protein	*

6.2 Appendix Table 2

List of positive & negative chromatin modifiers The genes from the clusters in Fig. 4.2, responsible for chromatin structure modification, and classified as enhancers and suppressors [20, 21] is shown. c+ - positive chromatin modifiers, DNA unwinding and enhance transcription c- - negative chromatin modifiers, suppress transcription

		H1
	Symbol	**Name**
c+	ARD1	ARD1 homolog, N-acetyltransferase (S. cerevisiae)
c+	PDX1	E3-binding protein
c+	EZH2	enhancer of zeste homolog 2 (Drosophila)
c+	HELSNF1	helicase with SNF2 domain 1
c+	HELLS	helicase, lymphoid-specific
c+	HMGA1	high mobility group AT-hook 1
c+	HMGA2	high mobility group AT-hook 2
c+	HMG20B	high-mobility group 20B
c+	HMGB3	high-mobility group box 3
c+	HMG2L1	high-mobility group protein 2-like 1
c+	HBOA	histone acetyltransferase
c+	SSRP1	structure specific recognition protein 1
c+	SUPT6H	suppressor of Ty 6 homolog (S. cerevisiae)
c+	SMARCA1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 1
c+	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4
c+	SMARCA5	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5
c+	SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1
c+	SMARCC1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1
c+	SMARCD1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1
c+	SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1
c-	DNMT1	DNA (cytosine-5-)-methyltransferase 1
c-	DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha
c-	DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta
c-	HDAC1	histone deacetylase 1
c-	HDAC2	histone deacetylase 2
c-	HDAC3	histone deacetylase 3
c-	BAT8	HLA-B associated transcript 8
c-	MBD3	methyl-CpG binding domain protein 3
c-	ORC2L	origin recognition complex, subunit 2-like (yeast)
c-	SAP18	sin3-associated polypeptide, 18kDa
c-	SIRT1	sirtuin (silent mating type information regulation 2 homolog) 1 (S. cerevisiae)
c-	SIRT3	sirtuin (silent mating type information regulation 2 homolog) 3 (S. cerevisiae)
		H2
c+	HMGN3	high mobility group nucleosomal binding domain 3
c+	HMG20A	high-mobility group 20A
c+	NCL	nucleolin
c+	SAFB	scaffold attachment factor B
c+	SMARCC2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2
c+	SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1

c-	CHD4	chromodomain helicase DNA binding protein 4
c-	HEMK	HEMK homolog 7kb
c-	MTA1	metastasis associated 1
c-	SAP18	sin3-associated polypeptide, 18kDa
c-	TGIF2	TGFB-induced factor 2 (TALE family homeobox)
		H3
c+	DEK	DEK oncogene (DNA binding)
c+	HMGN4	high mobility group nucleosomal binding domain 4
c+	SUPT4H1	suppressor of Ty 4 homolog 1 (S. cerevisiae)
c+	SMARCA1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 1
c+	TAZ	transcriptional co-activator with PDZ-binding motif (TAZ)
c-	MBD2	methyl-CpG binding domain protein 2
c-	SALL1	sal-like 1 (Drosophila)
		H4
c+	BAZ1A	bromodomain adjacent to zinc finger domain, 1A
c+	MORF	monocytic leukemia zinc finger protein-related factor
c+	NCOA2	nuclear receptor coactivator 2
c+	NCOA3	nuclear receptor coactivator 3
c+	RUNXBP2	runt-related transcription factor binding protein 2
c+	SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2
c_	MGMT	O-6-methylguanine-DNA methyltransferase
c-	BMI1	B lymphoma Mo-MLV insertion region (mouse)
c-	DMTF1	cyclin D binding myb-like transcription factor 1
c-	EZH1	enhancer of zeste homolog 1 (Drosophila)
c-	HRMT1L1	HMT1 hnRNP methyltransferase-like 1 (S. cerevisiae)
c-	MBD2	methyl-CpG binding domain protein 2
c-	MBD4	methyl-CpG binding domain protein 4
c-	SP100	nuclear antigen Sp100
c-	PFDN5	prefoldin 5
c-	SIRT2	sirtuin (silent mating type information regulation 2 homolog) 2 (S. cerevisiae)
c-	SHARP	SMART/HDAC1 associated repressor protein
c-	SP110	SP110 nuclear body protein
		H5
c+	BAZ1A	bromodomain adjacent to zinc finger domain, 1A
c+	PCAF	p300/CBP-associated factor
c-	HDAC4	histone deacetylase 4
c-	MBD4	methyl-CpG binding domain protein 4
c-	SP100	nuclear antigen Sp100

		K1
c+	BAZ1B	bromodomain adjacent to zinc finger domain, 1B
c+	BAZ2B	bromodomain adjacent to zinc finger domain, 2B
c+	SAS10	disrupter of silencing 10
c+	PDX1	E3-binding protein

c+	HMG20A	high-mobility group 20A
c+	HMGB1	high-mobility group box 1
c+	HMGB2	high-mobility group box 2
c+	HMGB3	high-mobility group box 3
c+	HMG2L1	high-mobility group protein 2-like 1
c+	HAT1	histone acetyltransferase 1
c+	SSRP1	structure specific recognition protein 1
c+	SMARCA1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 1
c+	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4
c+	SMARCA5	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5
c+	SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1
c+	SMARCC1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1
c+	SMARCD1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1
c+	SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1
c+	SMARCF1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily f, member 1
c-	DMTF1	cyclin D binding myb-like transcription factor 1
c-	DNMT1	DNA (cytosine-5-)-methyltransferase 1
c-	Eu-HMTase1	euchromatic histone methyltransferase 1
c-	HDAC2	histone deacetylase 2
c-	MTA1	metastasis associated 1
c-	MBD3	methyl-CpG binding domain protein 3
c-	MBD4	methyl-CpG binding domain protein 4
c-	NCOR1	nuclear receptor co-repressor 1
c-	SAP30	sin3-associated polypeptide, 30kDa
c-	SIRT3	sirtuin (silent mating type information regulation 2 homolog) 3 (S. cerevisiae)
c-	SUV39H1	suppressor of variegation 3-9 homolog 1 (Drosophila)
		K2
c+	DEK	DEK oncogene (DNA binding)
c+	HMGN4	high mobility group nucleosomal binding domain 4
c+	HMGB1	high-mobility group box 1
c+	HMGN2	high-mobility group nucleosomal binding domain 2
c+	ELP3	likely ortholog of mouse elongation protein 3 homolog (S. cerevisiae)
c+	NCL	nucleolin
c+	SMARCC1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1
c-	MBD4	methyl-CpG binding domain protein 4
c-	ORC2L	origin recognition complex, subunit 2-like (yeast)
c-	SIRT5	sirtuin (silent mating type information regulation 2 homolog) 5 (S. cerevisiae)
		K3
c+	ARD1	ARD1 homolog, N-acetyltransferase (S. cerevisiae)
c+	HELSNF1	helicase with SNF2 domain 1
c+	HELLS	helicase, lymphoid-specific
c+	HMGA1	high mobility group AT-hook 1
c+	HMG2L1	high-mobility group protein 2-like 1
c+	HBOA	histone acetyltransferase
c+	SUPT4H1	suppressor of Ty 4 homolog 1 (S. cerevisiae)
c+	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4
c+	SMARCD1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1

c+	SMARCF1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily f, member 1
c-	CHD4	chromodomain helicase DNA binding protein 4
c-	DNMT2	DNA (cytosine-5-)-methyltransferase 2
c-	DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha
c-	DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta
c-	HEMK	HEMK homolog 7kb
c-	HDAC1	histone deacetylase 1
c-	BAT8	HLA-B associated transcript 8
c-	SALL1	sal-like 1 (Drosophila)
c-	TGIF2	TGFB-induced factor 2 (TALE family homeobox)
		K4
c+	SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2
c-	CBX4	chromobox homolog 4 (Pc class homolog, Drosophila)
c-	KRT23	keratin 23 (histone deacetylase inducible)
c-	MBD2	methyl-CpG binding domain protein 2
c-	SET07	PR/SET domain containing protein 07
c-	SIRT2	sirtuin (silent mating type information regulation 2 homolog) 2 (S. cerevisiae)
c-	SIRT6	sirtuin (silent mating type information regulation 2 homolog) 6 (S. cerevisiae)
		K5
c+	PCAF	p300/CBP-associated factor
c-	SIRT7	sirtuin (silent mating type information regulation 2 homolog) 7 (S. cerevisiae)

פרופיל גנטי של תאי גזע אנושיים

מיכל גולן-משיח

תזה לדרגת מוסמך מוגש למועצה המדעית של מכון ויצמן למדע

בהדרכת

פרופסור איתן דומאני ופרופסור דוד גבעול

דצמבר 2003