Master Thesis

Gene Expression Analysis of Mesenchymal Stem Cell Differentiation and Leukemic Over Expression of Tissue Specific Genes

Dvir Netanely

May 2006

Eytan Domany's Group

Physics of Complex Systems

Weizmann Institute of Science

Rehovot, Israel

Acknowledgements

Firstly, I would like to thank Professor Eytan Domany, my supervisor, for making it all happen. He is definitely the one to thank for creating an environment that is both fun and challenging, he has exposed me to many interesting fields of research and I learned from him much more than he realizes. Eytan – you're one of a kind.

Assif Yitzhaky and Hilah Gal were my roomies; Assif taught me everything I know about Matlab and about working with gene expression and Hilah devotedly took care of my mental health. It was a pleasure sharing this time with them.

It was a great honor collaborating with Professor Leo Sachs and Dr. Joseph Lotem on the cancer project, and with Professor David Givol on the mesenchymal stem cell project. They have inspired me greatly and amazed me with their diligence and vast biological knowledge.

I wish to thank Prof. Dan Gazit and Dr. Hadi Haslan from the Hebrew University for the collaboration on the mesenchymal stem cell project.

I would also like to thank Eytan's talented past and present group members for helping me with my work and for making it a fun thing to do: Noam Shental, Hilah Benjamin, Shiri Margel, Michal Mashiach, Or Zuk, Roman Brinzanik, Liat Ein-Dor, Garold Fuks, Libi Hertzberg, Itai Kela, Anat Reiner, Jacob Bock Axelsen, Shlomo Urbach, Mark Koudritsky and Paz Polak. Special thanks to Michal Sheffer, Tal Shay, Yuval Tabach and Dafna Tsafrir – for sharing their experience with me and for teaching me so many things.

I also wish to thank Mr. Yossi Drier for the technical support.

Irit Fishel – my dear girlfriend shared the weight of giving birth to this thesis with me and was a source of endless wise insights and trustworthy statistical support.

Finally – to my beloved family – thank you guys for everything.

THANK YOU ALL AND GOOD LUCK IN ALL YOUR FUTURE ENDEAVORS!

Dvir

Contents

Abstract

The DNA microarray technology enables simultaneous measurement of the expression levels of thousands of genes in cells of a given biological sample. It provides a high-throughput quantitative survey of the transcriptional activity within the sample cells by measuring the mRNA concentration of many genes. In this work, we have used clustering algorithms and various statistical methods to analyze gene expression data in two different studies. In the first study, we have researched mesenchymal stem cell differentiation by analyzing 17 human samples including embryonic stem cells, mesenchymal stem cells and differentiated fat and bone samples. Our analysis explored general properties of the dataset and also identified different groups of genes involved in the differentiation process. The second study dealt with the identification of genes that are over expressed in human cancer and also show specific patterns of tissue-dependent expression in normal tissues. To this end, we have analyzed gene expression data from three different kinds of samples: normal human tissues, human cancer cell lines and leukemic cells from lymphoid or myeloid leukemia pediatric patients. The results indicate that many genes that are over expressed in human cancer cells are specific to a variety of normal tissues, including normal tissues other than those from which the cancer originated.

Part 1

General Methods

DNA microarray technology

The living cell is a dynamic system, continuously changing through developmental pathways and in response to environmental conditions. The cell changes its properties by producing different subsets of proteins at different times according to its functional needs. Out of the entire genomic repertoire, only needed genes are transcribed to messenger RNA (mRNA) molecules, which in turn are translated to sequences of amino acids composing the proteins. Gene expression regulation on the transcription level is one of the major known gene control mechanisms. It involves a complex network of transcription factors acting to activate or repress the expression of their target genes.

The set of genes transcribed in a cell (the cell *transcriptome*, representing the collection of all transcribed mRNAs floating in the cell) therefore reflects the current cell "state", and can tell us a lot about the genetic makeup, response environmental conditions and developmental stage of the examined cell. Clearly, this is an approximation, since a cell's "state" depends on a variety of other factors, such as protein concentration, chemical changes on the protein level (such as phosphorilation and complex formation), protein localization in the cells and more. Nevertheless, one assumes that knowledge of the transcriptome does provide a relevant characterization of the biological state of a cell (and tissue).

The DNA microarray technology enables simultaneous measurement of the expression levels of thousands of genes in cells of a given biological sample. It provides a high-throughput quantitative survey of the transcriptional activity within the sample cells by measuring the mRNA concentration of many genes.

DNA microarray technology is based on the tendency of a given mRNA molecule, extracted from cells in the experimental system, to specifically hybridize by base-pairing to a complementary DNA sequence located on the microarray.

There are several types of DNA microarray technologies; however, this work will focus on high-density oligonucleotide microarrays manufactured by Affymetrix, using their patented 'GeneChip' technology (For a review of the available microarray technologies, see [1]).

Figure 1. Affymetrix GeneChip microarray

Affymetrix' GeneChip microarray is a coated quartz surface divided to many thousands of cells forming a two dimensional array. Each microarray cell (named a *feature* by Affymetrix terminology) contains many short identical single-stranded DNA fragments (named *probes*), that are imbedded on the chip surface using photolithography during the microarray manufacturing process. The Affymetrix HG-U133A microarray, which is used in this work, contains 500,000 features of size 11 μm each. Each feature contains thousands of 25 nucleotide long DNA probes [2].

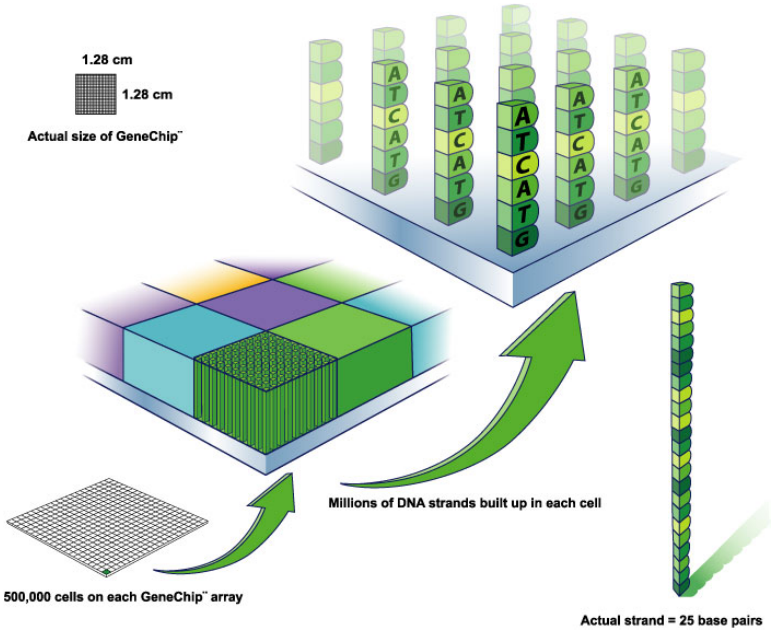


Figure 2. Affymetrix GeneChip microarray is composed of thousands of spots, each one containing millions of 25 base-pair long probes.

Microarrays are used to measure gene expression by extracting mRNA from the experimental sample (for example body tissue or cultured cells), converting it to complementary DNA (cDNA) which is easier to amplify than RNA, reverse transcribing it to RNA which is then fragmented and tagged with a fluorescent label that will enable to measure the hybridization level for each feature independently. The resulting solution is injected onto the microarray and so RNA fragments originated from the experimental sample are hybridized, with different affinities, to the probes on the array. The microarray is then washed and scanned with a laser scanner that yields a quantitative reading of the fluorescent light. The fluorescencnt light intensity of each microarray feature is proportional to the number of RNA molecules that hybridized to the feature's probes.

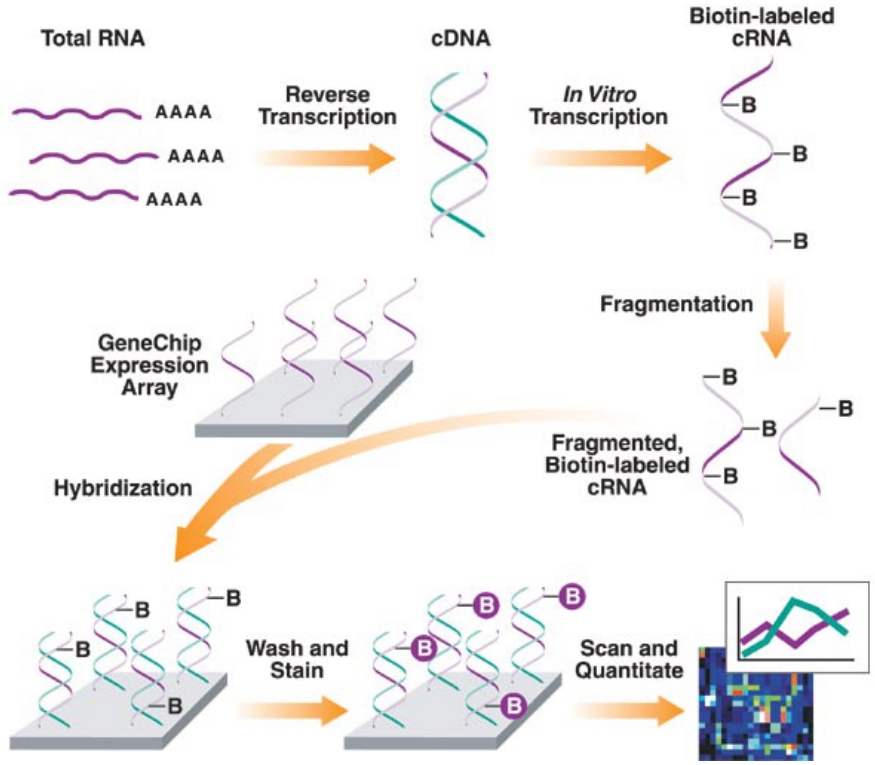


Figure 3. Standard eukaryotic gene expression assay. Labeled cDNA or cRNA targets derived from the mRNA of an experimental sample are hybridized to nucleic acid probes attached to the solid support. By monitoring the amount of label associated with each DNA location, it is possible to infer the abundance of each mRNA species represented.

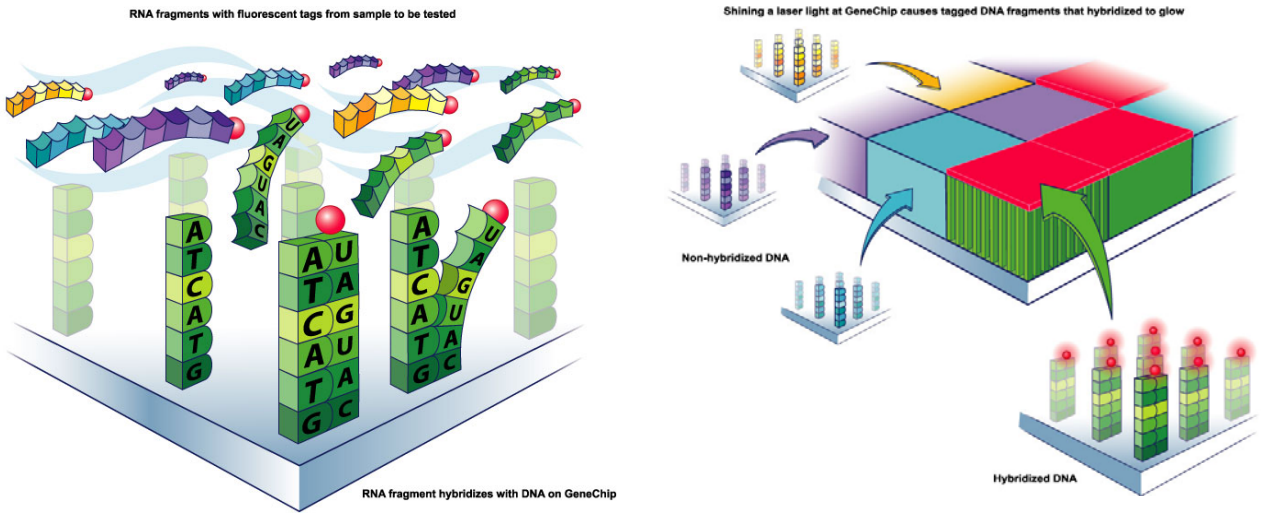


Figure 4. Hybridization of fluorescently tagged mRNA sample to the microarray probes

The HG-U133Av2 microarray is capable of measuring the expression levels of more than 14,500 genes represented by 22,788 different probe-sets.

The expression of each gene is measured based on the hybridization of RNA extracted from an experiment sample (called *target*) with several probe pairs located on the microarray. Each probe pair consists of Perfect-Match probe (PM) and a Mis-Match probe (MM); The Perfect Match probe is a 25 base long oligonucleotide, which is a perfect complement to a 25 base long sub-sequence of the target gene. The Mismatch probe differs from the Perfect-Match probe in one nucleotide, positioned in the middle of the probe. It is used for specificity control, enabling evaluation and subtraction of background noise and unspecific hybridization.

The HG-U133Av2 microarray contains 11 probe pairs for each target gene; this group of probes, aimed at capturing a specific transcript, is called a Probe-Set. Using several independent probe pairs (instead of just one pair) to detect the concentration of a certain RNA molecule, significantly increases the measurement accuracy. Probe-sets are designed, using the known genomic sequences, to be as specific as possible for the target gene sequence, reducing false positives and miscalls[3].

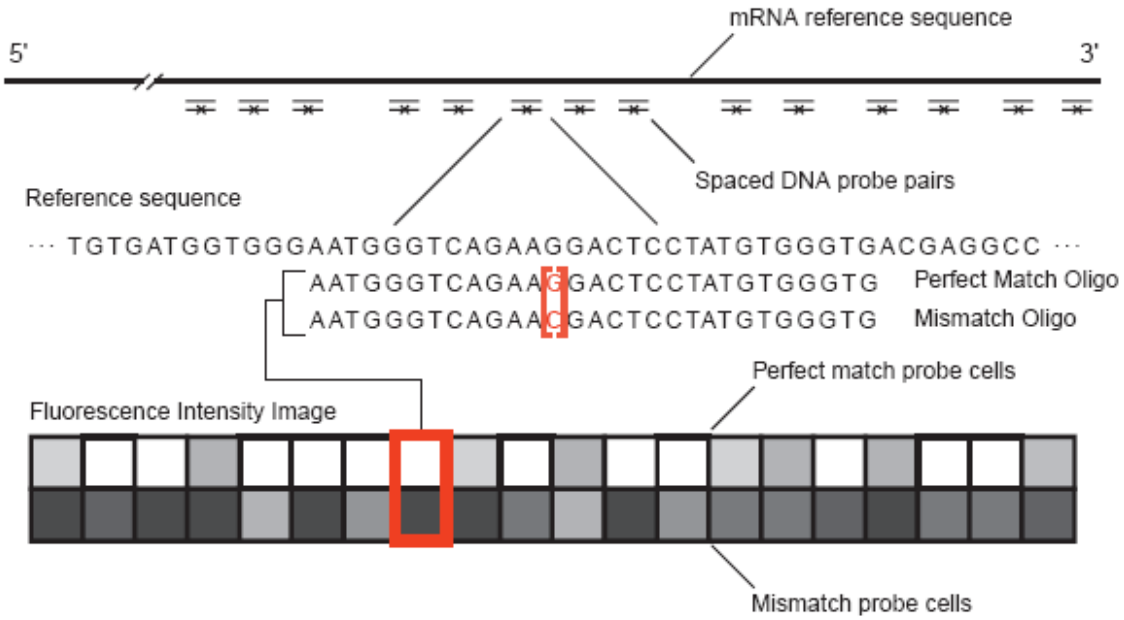


Figure 5. Oligonucleotide probes and the probe-set. Probes are 25 base long oligonucleotide sequences chosen from RNA reference sequence. The probe set contains 11 pairs of PM and MM probe cells. Each probe cell contains millions of copied of the cell-specific oligonucleotide probe.

It is also worth mentioning that microarrays usually contain more than one probe-set per gene, thus enabling to distinguish different transcript isoforms generated due to alternative splicing or other mechanisms.

After scanning the microarray, the fluorescence intensity for each probe is stored. The final expression measurement for each given gene is calculated as a weighted average of all probe pairs representing the gene, and can be conducted in several ways – each having its advantages and disadvantages.

In MAS 4.0, gene expression was calculated as the average of differences between the perfect-match and the mis-match probes of all the pairs representing the gene.

$$E = \frac{1}{n}\sum_{i=1}^{n}(PM_i - MM_i)$$

E is the expression value of one probeset, representing a certain gene. In our case, n=11 as probe-sets in the HG-U133Av2 microarray are composed of 11 probe-pairs.

According to the more recent MAS 5.0 algorithm, gene expression is calculated based on a similar principle (averaging of PM/MM differences), but in a more robust manner. It is using the one-step Tukey's biweight algorithm, which is a method to determine a robust average unaffected by outliers. For details, see http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

The MAS 5.0 algorithm also provides p-Values, calculated for each expression reading, representing its detection reliability and thus enabling removal of genes that were found "absent" on all or on most of the samples from the dataset.

Dataset compilation and preprocessing

General

A typical microarray experiment is aimed at identifying differences in gene expression between two or more biological conditions such as tissue samples taken from healthy individuals versus cancer patients, samples taken from different cell lines, or samples taken at different time points along some biological pathway. In addition, a good microarray experimental design should include sample replications - ideally biological independent replications (enabling to asses the biological variation), rather than technical replications (using same biological sample on multiple arrays, used to asses measurement variability)[4].

A typical gene expression analysis therefore involves working with data originating from a set of microarrays – a gene expression *dataset*. Datasets originating from a microarray experiment composed of as many samples as possible enables the employment of powerful statistical methods to detect genes that are differentially expressed in one sample group compared to the other, and optimizes the potential of the dataset-based analysis to yield solid reliable results.

After all microarrays of the experiment are produced and scanned, their data is joined to compose one big table that will be the basis for gene expression analysis. In this table, each column contains data coming from one microarray (sample), and each row represents a probe-set (gene). See, for example, Table 1 on the next page.

Probeset ID	Gene Symbol	S1	S1 detection	S2	S2 detection	S3	S3 detection	S4	S4 detection
1316_at	THRA	16.8	A	11.7	A	21.9	A	26.3	P
1320_at	MMP14	20.3	A	27	A	17	A	16.9	A
1405_i_at	TRADD	15	P	8.3	A	2.2	A	0.5	A
1431_at	FNTB	4.8	A	7.5	A	6.4	A	8.6	A
1438_at	PLD1	16.7	P	3.9	A	4.5	A	26.5	P
1487_at	PMS2L11	106.9	P	155.6	P	82.8	M	68.5	A
1494_f_at	BAD	12.8	A	12	A	12.7	A	2.4	A
1598_g_at	PRPF8	4532.9	P	4103.1	P	3302.4	P	2831.4	P
160020_at	CAPNS1	284.8	P	271	P	288.4	P	316.8	P
1729_at	RPL35	135.6	P	148.3	P	129.5	P	121.2	P
1773_at	RPL28	61.2	P	50.1	A	55.5	P	45.5	P
177_at	MMP14	11.9	P	16.1	P	11.6	A	16.8	P
179_at	TRADD	53.1	A	33.3	A	57	A	24.6	A
1861_at	FNTB	127.7	P	128.6	P	117.1	P	141.4	P
200000_s_at	PLD1	577.6	P	534.7	P	492.5	P	517.3	P
200001_at	PMS2L11	2045.8	P	2277.2	P	1635.1	P	1837.6	P
200002_at	BAD	5472.2	P	5159.8	P	4648.6	P	4342.6	P
200003_s_at	PRPF8	7850.8	P	6521.3	P	5837.3	P	7389.4	P

Table 1. Example dataset table. The above table contains 18 rows, each representing a probe-set (associated with a gene symbol). The table contains data coming from 4 microarrays, measuring gene expression in sample S1, S2, S3 and S4. The first column for each sample indicate the expression signal, whereas the second one contain the detection call – **A**bsent or **P**resent.

Several standard steps are routinely conducted before the data can be successfully and efficiently analyzed. The term 'preprocessing' is used to describe a series of mathematical manipulations conducted on the data, making it compatible to the subsequent high-level analyses.

Preprocessing goals usually include the following:

1) Reducing dataset dimensionality by removing un-informative probe-sets, such as probe-sets exhibiting low variability over the samples.
2) Applying mathematical transformations that will moderate the effect of outliers and emphasize mid-range expression values.

Different analyses may require different preprocessing steps. Here are some of the most commonly used preprocessing steps applied to a standard gene expression dataset:

Scaling

Scaling is a transformation of the expression values conducted on the array level, aimed at making data originating from different microarrays comparable. In this work, we have used the scaling conducted by the MAS 5.0 algorithm, which applies scaling on each microarray independently, by bringing the average of all expression values spanning between the 2^{nd} and 98^{th} percentile in the analyzed microarray to a predefined target average (usually set to ~250). The key assumption of the global scaling strategy is that most of the genes do not change between the analyzed arrays, and therefore the values of each microarray should roughly have the same average.

Removal of all-absent probe-sets

Affymetrix MAS 5.0 algorithm provides each probe-set reading in a given microarray with a *detection* p-value. The detection p-value indicates whether the corresponding transcript is reliably detected (Present) or not detected (Absent). Calculation of the detection p-value is based on probe pair intensities, and is then compared against a user-defined cutoff (of 0.05 in most cases) to be translated to the Absent/Marginal/Present detection calls. (For more details, please refer to http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf)

A standard microarray data preprocessing step is to remove probe-sets that are labeled as 'absent' on all (or most) dataset samples. Such genes are of little interest, as they were not detected as expressed on any of the experiment samples and thus do not seem to play a role in the investigated biological process.

Applying Log2 transformation

A common early step in microarray data analysis is log transformation. Many statistical methods are based on the assumption that measurement errors are additive and hence normally distributed. In microarray data there is evidence that indicates that the errors are multiplicative. Hence applying Log2 transformation brings the noise distribution close to the normal distribution and, in addition, quenches the data to reduce the effect of outliers [5, 6].

Setting a threshold

Data generated by microarrays is known to be noisy in the low value range due to measurement noise. In the following figure, a scatter plot of log2 transformed expression data of two microarray replicates is shown (Same biological sample). The left figure displays the log2-transformed data before applying any threshold. Such a plot can help us determine the threshold value by identifying the minimal value at which the relation between the two replicates becomes linear. In the displayed figure, a reasonable threshold can be defined as ~3.

The right figure displays the data after applying a threshold of 3. Threshold is applied by setting to the threshold, all expression values below it.

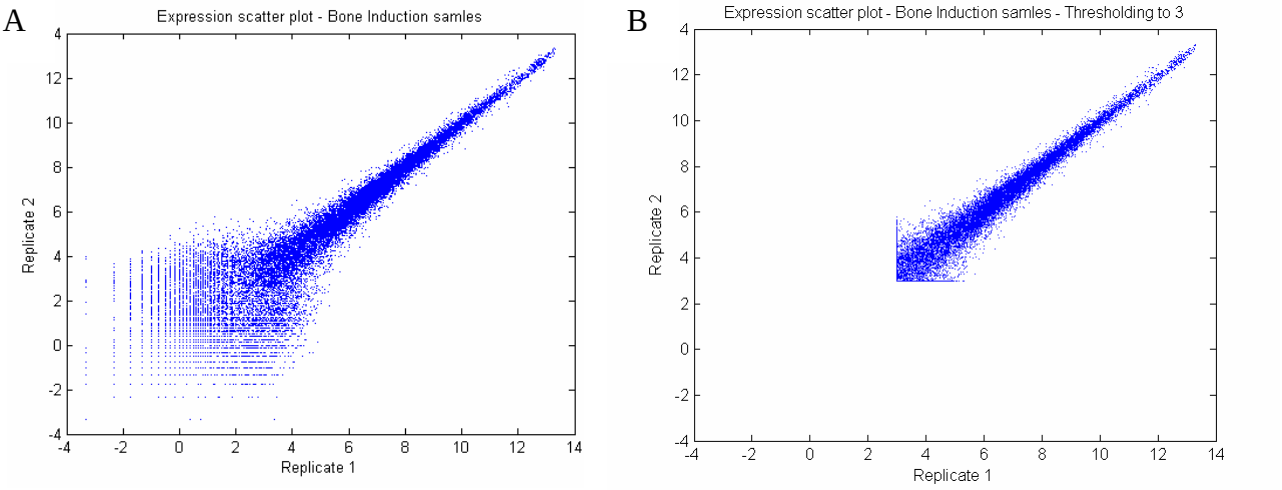


Figure 6. Scatter plots of replicates expression before setting a threshold of 3 (A) and after (B).

Variability filter

Since we are usually interested in genes whose expression changes between the experiment samples, a typical preprocessing procedure includes the removal of genes whose expression variance is below a certain variability threshold. This threshold is usually defined depending on the number of probe-sets in our capacity to computationally process later on.

Assuming a given dataset includes n_s samples (columns) and n_g probe-sets (rows), and that E_{gs} represents the expression value of gene g on sample s, the standard deviation of probe-set *g* (representing a gene) is denoted as

$$\sigma_g = \sqrt{\frac{\sum_s (E_{gs} - \overline{E_g})^2}{n_s - 1}}$$

Centering and Normalization

Quite often, we are interested in the way genes *relatively* change their expression between samples rather than absolutely. Therefore, before applying high-level analysis on the data, we first *standardize* the dataset rows (each row corresponds to a probe-set/gene). Standardization includes centering (mean of each gene equals 0) and normalization (standard deviation of each gene equals 1).

The following equation demonstrates how standardization is performed;

$$E'_{gs} = \frac{E_{gs} - \overline{E}_g}{\sigma_g}$$

Comments

Through this work, the terms 'probe-sets' and 'genes' are used interchangeably.

'Sample types' are used interchangeably with 'sample groups'.

Unless stated otherwise, rows represent genes and columns represent microarrays or samples.

Supervised data analysis methods

Analysis of gene expression data is a challenging task due to the high dimensionality of a typical microarray dataset. Many statistical methods and bioinformatic algorithms have been developed or adopted from other research fields in order to face this challenge. In general, these methods are aimed at identifying statistically significant expression patterns that are inherent in the usually noisy expression data. This section includes a brief overview of several data analysis techniques that were used in this work.

We start with a group of *supervised* methods, characterized by the use of external labels (such as clinical label of the dataset samples or functional class of genes). These methods are routinely used to identify genes that are differentially expressed in two or more sample subsets representing different biological conditions.

Fold change

'Fold Change' is a metric for comparing gene expression levels between two distinct experimental conditions. It is one of the first methods used to identify differentially expressed genes and it is still very popular today. 'Fold change' represents the ratio between the averaged expressions of a given gene in one sample group versus its averaged expression in a second sample group. For log transformed data, fold change is calculated (for each gene independently) as the difference between the means (or medians) of the two sample groups. Nowadays, fold change is applied mainly as a measure of effect size, used to rank genes by their expression difference between two sample groups. It is considered to be an inadequate inference statistic because it does not incorporate variance and offers no associated level of 'confidence' [4].

T-test and Rank-sum

The Student T-test and Mann-Whitney-Wilcoxon Ranksum test are statistical hypothesis tests used to assess whether the means of two groups are statistically different from each other [7, 8].

An important and common question in microarray experiments is the identification of differentially expressed genes between two distinct groups of samples (e.g. genes that are differentially expressed in normal versus tumor tissue). The basic statistical approach is to test for each gene the null hypothesis by which the gene is similarly expressed between the two groups (test for equality of means). If the P value that is calculated is less than the threshold chosen for statistical significance (usually the 0.05 level), then the null hypothesis that the two groups do not differ is rejected in favor of the alternative hypothesis, which typically states that the groups do differ.

T-test is a parametric test; it assumes that the data is normally distributed. The Rank-sum test is a non-parametric test used when the data is not normally distributed. The Rank-sum test is thus more permissible in its requirements, but the trade off is a reduced statistical power.

ANOVA

Like the t-test and the rank-sum test, ANOVA (Analysis Of Variance) is a statistical test used to assess means equality between groups; however, ANOVA is used to compare the means of more than two groups.

ANOVA tests for mean differences between groups by analyzing the variance, that is, by partitioning the total variance into the component that is due to true random error (variance within groups) and the components that are due to differences between means. These variance components are then tested for statistical significance, and if significant, we reject the null hypothesis of no differences between means, and accept the alternative hypothesis that the means (in the population) are different from each other [7, 8].

The multiplicity problem and FDR

The simultaneous testing of the null hypothesis in many thousands of genes in a DNA microarray dataset raises the *multiplicity problem*. The multiplicity problem refers to the situation where the expected number of `false discoveries` becomes large relative to the number of true discoveries. For example, if we use the customarily statistical threshold of α=0.05 on a microarray experiment of 10,000 genes, where 50 genes are truly differentially expressed, then we can expect approximately (10,000-50)*0.05 ~ 500 false positives (genes that are not truly differentially expressed but did pass the independent statistical tests).

The multiplicity problem was originally addressed by methods to control the family-wise type I error rate (FWER) which is the probability of having at least one false significant test result within the set of tested hypotheses. The simplest FWER approach is the 'Bonferroni correction' method. This method controls the group-wise error rate by rejecting the null hypothesis for a threshold of $\alpha' = \frac{\alpha}{N}$ where N is the number of tests performed. The division of the test-wise significance level by the number of tests insures that the expectancy of false positives is α, and thus the probability to get even one false positive is less or equal to α. A major drawback of this method is that it is too conservative. When the number of tests is high, such as in microarray experiments, legitimately significant results will fail to be detected.

Recently, Benjamini and Hochberg [9] have proposed a less conservative approach to multiple testing which calls for controlling the expected proportion of falsely discovered predictions among the list of predictions that are identified; the expected proportion is called the false discovery rate (FDR).

Let R denote the number of hypotheses rejected by the procedure, V the number of true null hypotheses that are wrongly rejected. Then:

$$FDR = E\left(\frac{V}{R}\right)$$

For example, if the FDR procedure returns 100 genes with a false discovery rate of 0.25 then we should expect 75 of them to be correct.

Gene Ontology (GO) and gene class testing

Gene ontology (GO) is a gene annotation system which is based on a hierarchical vocabulary that is species-independent. GO is used for describing gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. 'Biological Process' refers to biological goal or objective (example 'biological process' terms include *mitosis*, *DNA replication* or *metabolism*). 'Molecular Function' refers to the biochemical activity of the gene product (i.e. *DNA binding*, *ATPase activity*). Lastly, the 'Cellular Component' GO category refers to the location or complex of the given gene product (i.e. *nucleus*, *cell-membrane*). Each gene is independently assigned with GO terms from any of the 3 GO categories, and usually with more than one term from each category [10].

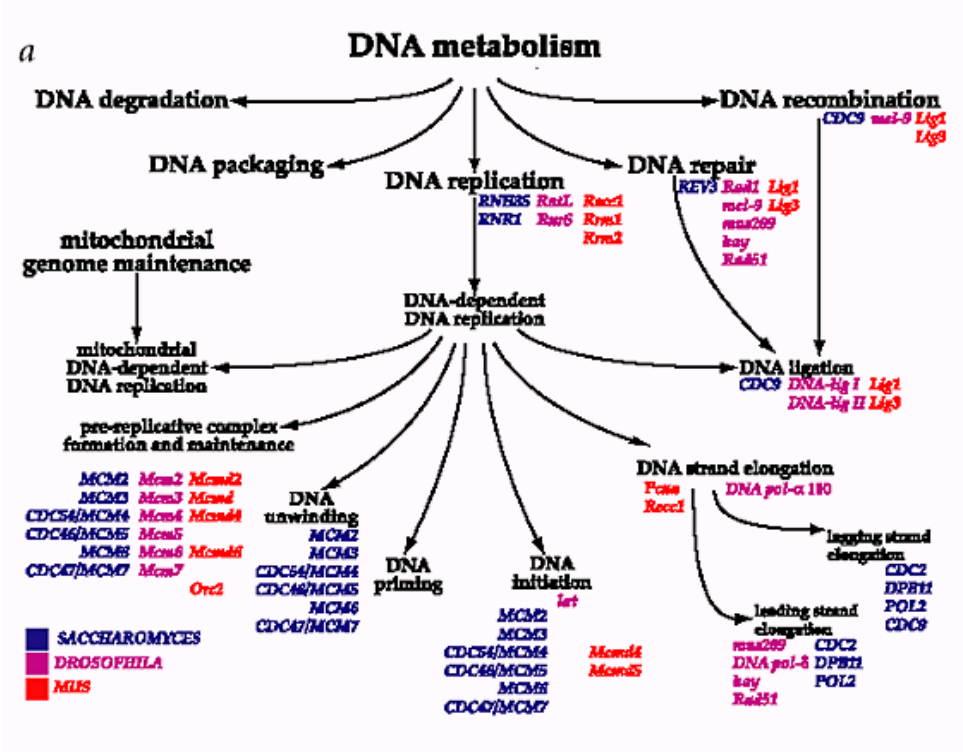


Figure 7. Example: The 'DNA metabolism' GO term and its descendant terms, which are part of the 'biological process' class.

The result of analyzing microarray datasets is often a list of differentially expressed genes or a list of genes included in a given gene-clusters. In an attempt to interpret such gene lists, they are analyzed in terms of the functional categories of the genes – usually based on Gene Ontology (GO) categories. A set of genes which is found to be enriched with a certain GO term, is more likely to be involved in the underlying biological process. GO enrichment is routinely used to validate the "biological sense" of a given set of genes.

Gene class testing identifies functional GO categories over-represented in a gene list relative to the representation within the proteome of a given species. Hyper-geometric based p-value is calculated for each GO term, assessing its over-representation in a given cluster compared with the total number of probe-sets on the microarray.

Gene class testing conducted in this work is based on GO annotations downloaded from the Affymetrix web site, updated to January 2006. Enrichment analysis was conducted using the *Profiler* software.

Unsupervised data analysis methods

Unsupervised methods use only the expression data points for the analysis without relying on any external predefined data labels. Clustering algorithms are an implementation of unsupervised learning approach, and they are used to organize huge numbers of unlabeled data points in a gene expression dataset into a structure. Each cluster within that structure contains a collection of data points that are similar to each other and have a similar expression pattern. Clustering can be applied on genes (rows) as well as on samples (columns). Gene clustering is used to explore gene expression assuming that genes whose expression is correlated (and thus are assigned to the same gene cluster), may have a related function. Examination of the produced clusters may provide insights on different biological processes reflected in the data.

Hierarchical clustering

In Hierarchical clustering [11], the expression data is partitioned to clusters in a series of steps. The algorithm iteratively joins the two closest clusters starting from singleton clusters (agglomerative hierarchical clustering) or iteratively partitioning clusters starting with the complete set (divisive hierarchical clustering). After each joining of two clusters, the distances between all the other clusters and the new joined cluster are recalculated. The complete linkage, average linkage, and single linkage methods use maximum, average, and minimum distances between the members of two clusters respectively. Like several other clustering algorithms, hierarchical clustering may be represented by a two dimensional diagram known as dendrogram, which illustrates the fusions or divisions made at each successive stage of analysis. Note that for hierarchical clustering, in order to obtain a particular partitioning into clusters, the distance metric, linkage methods and threshold distance must be defined by the user.

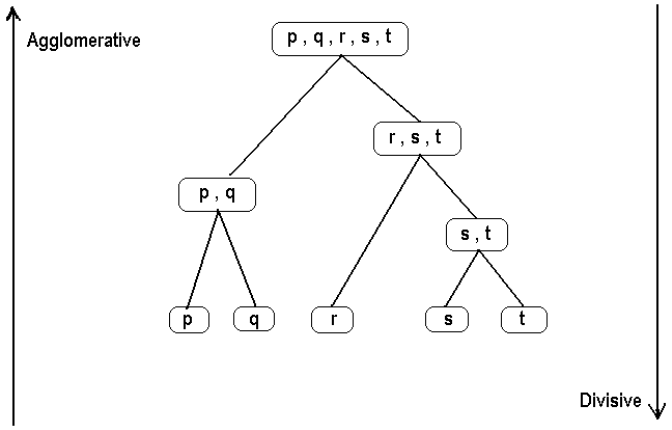


Figure 8. Example of a dendrogram representing hierarchical clustering.

The SPC clustering algorithm

The Super Paramagnetic Clustering algorithm [12] is based on the properties of an inhomogeneous ferromagnetic model. SPC is used to yield a temperature dependant hierarchical clustering of the given data (higher temperature values yield a higher resolution clustering, where at very high temperatures each data point is assigned to a different cluster). 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 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, by averaging over all 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.

SPC's advantages over other clustering algorithms include robustness against noise, creation of a hierarchy based clustering represented by a dendrogram. Furthermore, SPC does not require the specification of the number of clusters in

advance; SPC provides a reliable cluster stability measure that is used to define final output clusters. SPC uses a Euclidean distance measure.

CTWC

The Coupled Two Way Clustering algorithm [13-15] is using iterative clustering executions in order to identify stable gene and sample clusters. The algorithm finds stable gene clusters using an external clustering algorithm (such as SPC), and then uses these clusters to find stable sample clusters. These sample clusters are again used to find stable gene clusters, and so on – until no additional stable clusters are found. On each such iteration, one subgroup is in focus, and therefore it is minimally affected by the noise present in the total dataset containing thousands of data points. In this work, CTWC was used as an envelope for SPC only and was not executed iteratively.

Gene expression profiling

In gene expression profiling (a method developed as part of this study), dataset samples are grouped by sample type and their expression values are averaged independently for each gene. The distribution of the averaged expression values is sliced to N bins, and each expression value is then mapped to one of the bins. The expression of every gene is then represented by a vector with an alphabet of size N (such as [1 2 1 5 5]).

Using this simplified representation of the expression data, genes sharing the same profile are clustered together. The output of the profiling operation is a set of gene clusters, each one with a defined expression profile.

N – The number of bins, is a user-defined parameter defining the resolution of the profiling operation.

Profiling is useful for several reasons:

- It is intuitive and simple to understand.
- Runs very fast and thus can be applied on a very large number of probe-sets.

- After profiling is applied on a given dataset, profiles can be filtered based on various criteria (such as 'all monotonically increasing profiles', or 'all profiles exhibiting minimal expression on sample type X') to form meta-profiles of interests.
- Profiling can be applied on both un-standardized data and standardized data.

The following figures demonstrate applying gene expression profiling on a sample dataset.

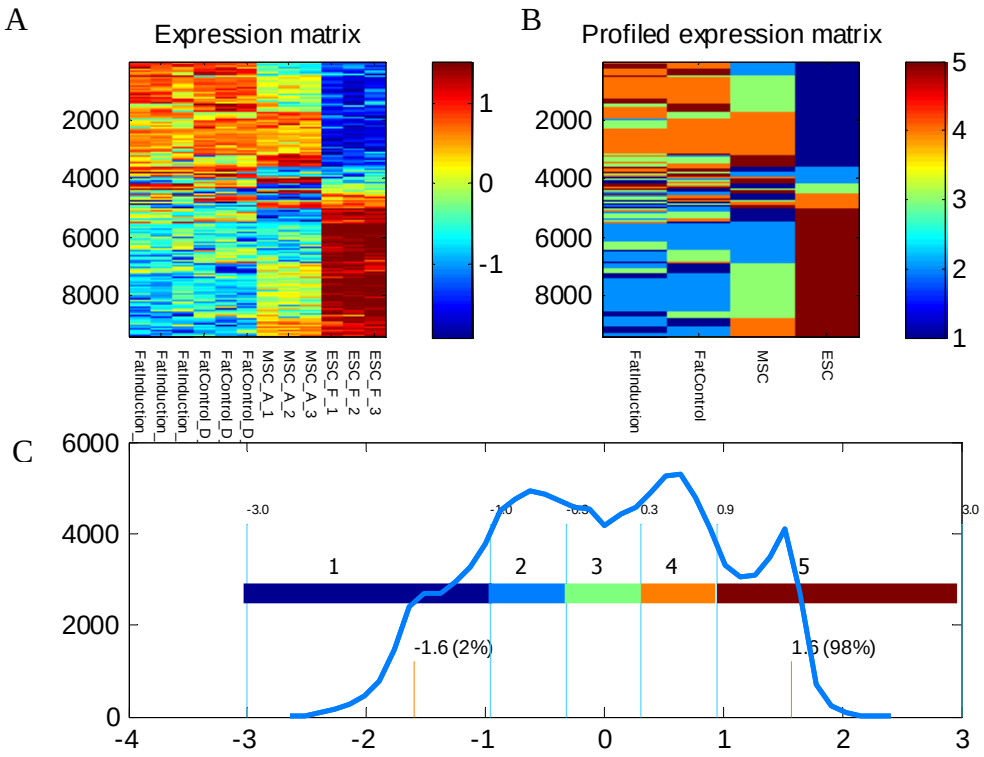


Figure 9. Gene Expression Profiling applied to sample dataset using a resolution of 5. In Profiling, expression is "simplified" by mapping expression values to several levels of expression. (A) Expression matrix of original sample dataset. Rows (representing genes) are ordered by profiles. (B) Profiled expression matrix. Each sample group is averaged and mapped to one of 5 expression level. (C) Distribution of dataset expression values, sliced to 5 intervals which defines the range of expression values mapped to each bin.

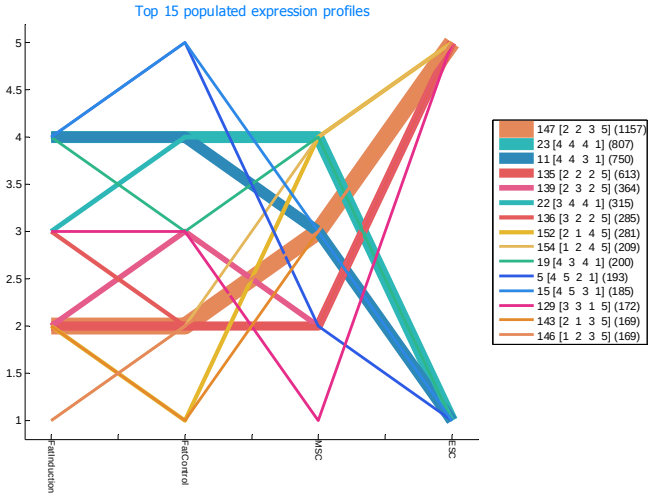


Figure 10. Top 15 populated expression profiles. The profiling operation conducted above yielded many different profiles. This figure displays the 15 largest profiles (containing the largest number of probesets). For example, profile #147 shown in orange, exhibits a profile of [2 2 3 5]: its probesets are expressed at low levels on the two left most samples types, and expressed at the highest level on the right most sample type. 1157 probe-sets are mapped to this profile, making it the most populated profile.

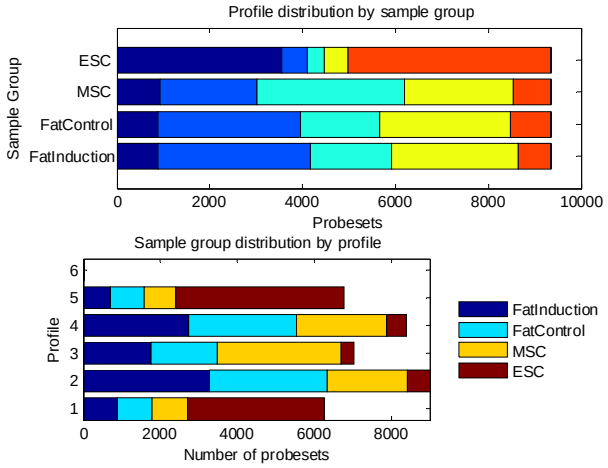
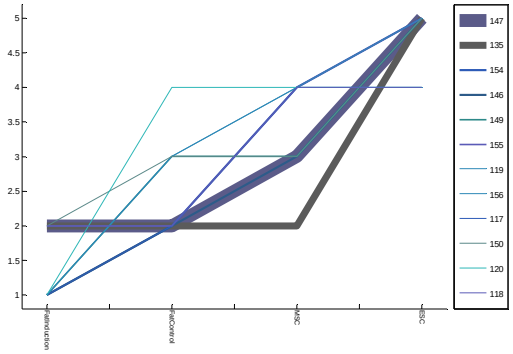


Figure 11. Analysis of profile distribution. Applying profiling on gene expression data, enables to analyze profile distributions. On the upper histogram we can see that the ESC sample group mainly express genes of level1 and of level 5. Looking on it from the opposite angle, on the lower histogram we can see that expression levels 1 and 5 are mainly occupied by the ESC sample group.

A



B

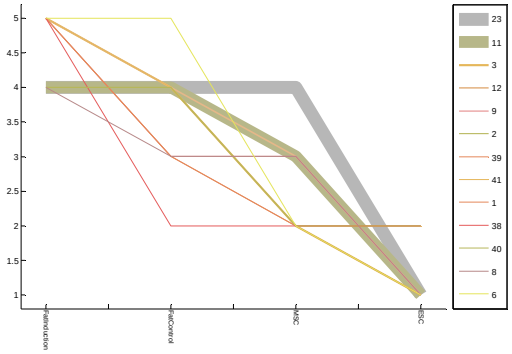


Figure 12. Filtering of gene expression profiles. Profiling gene expression data also enables to filter the profiles (and indirectly the genes that contain) according to different criteria. In this case we have looked for monotonically increasing (A) or monotonically decreasing (B) profiles that change their expression gradually along a certain biological process.

Part 2

Mesenchymal Stem Cell Differentiation

Biological Background

Stem cells are special kind of undifferentiated cells that can give rise to different types of mature cells [16]. Their main characteristics are multipotency, self-renewal and immortality. *Multipotency* refers to the ability of these undifferentiated cells to give rise to different types of mature cells. Their capacity for *self-renewal* enables them to proliferate and maintain their own cell population size. *Immortality* means that these cells do not die after a predetermined number of divisions.

There are several different types of stem cells, which differ in their differentiation potential (the range of mature cells they can differentiate into): The *totipotent* zygote, the *pluripotent* embryonic stem cells and the *multipotent* adult stem cells.

The *totipotent* **zygote**, formed by the fusion of an egg and a sperm cell upon fertilization, is the most potent stem cell of all. It has the capacity to generate an entire mammalian fetus and its surrounding supporting tissues. Within several days, the zygote develops into a blastocyst. The blastocyst is composed of a hollow ball of cells (Trophoblast) that will form the placenta, and a compact body of cells called inner cell mass (ICM), from which the fetus develops. The totipotent nature of the zygote is defined by its capacity to specialize into both the trophoblast and the ICM. The cells composing the ICM, develop to the 3 embryonic germ layers (ectoderm, mesoderm and endoderm) that will eventually give rise to the more than 200 mature differentiated cell types found in a mammalian organism [17].

Embryonic stem cells are cells derived from the inner cell mass of a 4-5 days old embryo that was created by in-vitro fertilization. Embryonic stem cells (ESCs)

are called *pluripotent* as their differentiation potential includes all three fetus germ layers that will differentiate during embryonic development into the more than 200 different mature cell types composing the adult organism. However, ESCs cannot differentiate into trophoblast (the extra-embryonic placenta progenitor) to form a complete blastocyst as the totipotent zygote can.

Murine embryonic stem cells were first isolated in 1981 [18], and human embryonic stem cells were isolated in 1998 [19]. Both exhibit normal and stable karyotype, express embryonic cell surface markers and can be cultured in vitro for very long periods in an undifferentiated state and yet retain their pluripotent differentiation potential.

In order to maintain their self-renewal and multi-lineage differentiation potential, both mouse and human embryonic stem cells were originally co-cultured in the presence of mouse embryonic fibroblast feeder layer that derives substances that block differentiation. Without a layer of feeder cells, cultured embryonic stem cells maintain their pluripotency only for a short time [20]. The feeder layer also provides the ESCs a sticky surface to which they can attach, and releases nutrients into the culture medium.

For mouse ESCs, it has been shown that continuous presence of leukemia inhibitor factor (LIF, a member of the interleukin-6 cytokine family) is sufficient to sustain self-renewal and pluripotency. LIF binds to the gp130 receptor on

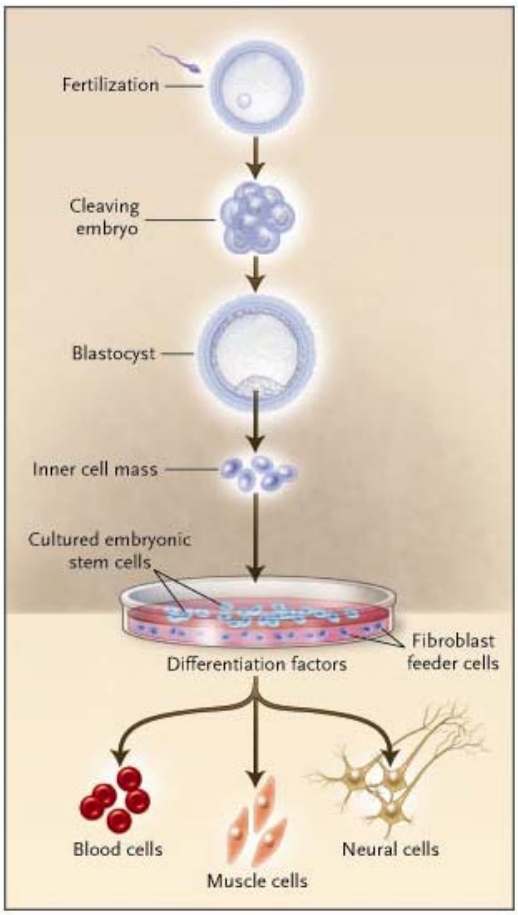


Figure 1. Derivation of Embryonic Stem Cells. Embryonic stem cells are derived from the inner cell mass of the blastocyst; cells composing the inner cell mass are isolated and then plated on culture medium, below which is a layer of feeder cells.

the murine ESC surface, which results in JAK kinase-mediated activation of the transcription factor STAT3 [21].

Human ESCs are indifferent to LIF, and it is not known to date which of the compounds derived from the fibroblast feeder cell layer (either of mouse origin, or of the more recently developed human fibroblast feeder cell layer) are responsible for keeping the cultured cells in an undifferentiated state.

Several molecular markers for undifferentiated pluripotent human ESCs have been identified. These markers are expressed in undifferentiated human ESCs and are turned off after differentiation. The identified markers Oct-4, Nanog, Rex1, TDGF1, Sox2, LeftyA, FGF4 are some of the most prominent [22],[23],[24]. Human ESCs also express high levels of telomerase [19]

Upon induction by specific differentiation compounds, cultured embryonic stem cells can differentiate in-vitro into a variety of mature cell types, including: neurons and skin cells (indicating ectodermal differentiation); blood, muscle, cartilage, endothelial cells, and cardiac cells (indicating mesodermal differentiation); and pancreatic cells (indicating endodermal differentiation) [22]. One of the most important goals of current stem cell research is the development of specific protocols for efficient directed differentiation of ESCs into any mature cell of interest.

Adult stem cells are *multipotent* stem cells found in the adult organism. Like Embryonic stem cells, they are capable of self-renewal throughout the organism's life, and also capable of differentiating into different mature cell types (usually through an intermediate cell of increased commitment called a progenitor). However, adult stem cells are already committed to a certain cell lineage and thus they are restricted in their differentiation range.

Adult stem cells reside within mature tissues and serve as a limitless source for new mature cells, enabling maintenance and repair of the tissue by continuously regenerating mature tissues either as part of normal physiology or as part of repair after injury.

Adult stem cells have been identified in many animal and human tissues, including blood, brain, skin, gut, muscle and in the mesenchyme – which is the focus of this work (see table 1) [23].

Table 1 | **Mammalian adult stem cells**

Tissue	Stem cell	Niche	Progeny
Blood	Haematopoietic stem cell	Endosteal surface of bone marrow	All myeloid and lymphoid blood lineages
Mesenchyme	Mesenchymal stem cell	Within bone marrow cavity	Bone, cartilage, tendon, smooth muscle, adipose tissue and stroma
Brain	Neural stem cell/ neurosphere	Subventricular zone and hippocampus	Neurons, glial cells and oligodendrocytes
Gut	Crypt cell/gut epithelial progenitor	Gut crypt	Enterocytes, enteroendocrine cells, goblet cells and Paneth cells
Heart	Cardiac progenitor	Not determined	Cardiac myocytes
Liver	Oval cell	Terminal biliary ductule	Hepatocytes and cholangiocytes
Pancreas	Pancreas-derived multipotent precursors	Not determined	Pancreatic endocrine and acinar cells
Skeletal muscle	Satellite cell	Between sarcolemma and basil lamina	Myocytes/myofibrils
Skin/hair	Bulge cell	Bulge in hair follicle	Epidermis, hair follicles and sebaceous glands
Male germ cells	A_s spermatogonia	Basement membrane of seminiferous tubule	Sperm

Table 1. Adult stem cells. Adult stem cells have been found in small amounts in many mature tissues.

Adult stem cells are usually found within compartments (called niches), where they respond to a variety of extrinsic signals that determine their fate. The stem cell niche is a dynamic multi-cellular structure, which serves as a controlled microenvironment, balancing the stem cell tendency to proliferate or to give rise to differentiated tissue cells. The exact interactions composing the microenvironments of the different stem cell niches are still mostly unknown [25]. Stem cells are considered quite rare, composing only a small fraction of the tissue cellularity.

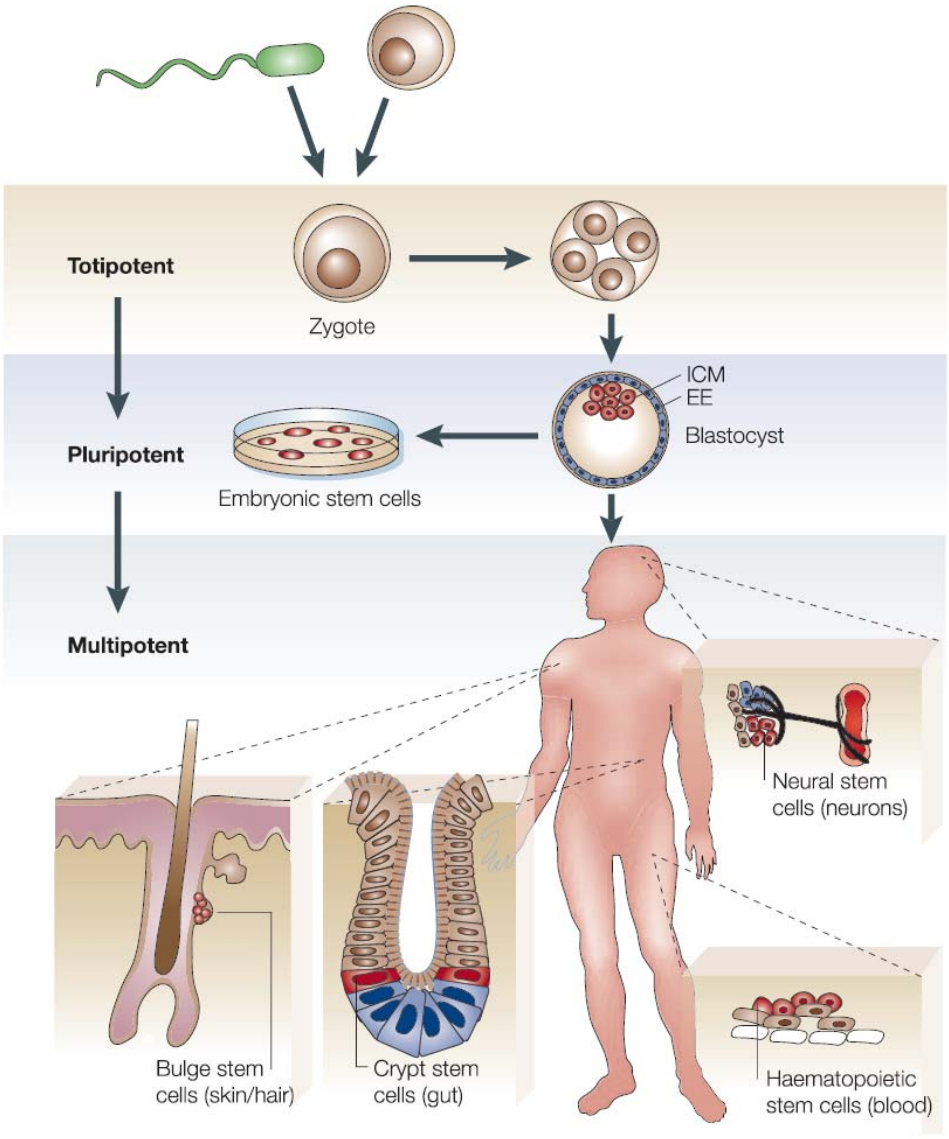


Figure 2. The hierarchy of stem cells. The totipotent zygote give rise to the blastocyst. Pluripotent embryonic stem cells derived from the inner cell mass of the blastocyst can be cultures in vitro. Multipotent adult stem cells exist in many mature tissues, used as a reservoir of renewing cells.

In recent years, an increasing body of research suggests that multipotent adult stem cells are much more flexible in their differentiation potential, capable of trans-differentiating across tissue lineage boundaries into mature cell types other than their tissue of origin. One example for adult stem cell plasticity is demonstrated by studies showing that hematopoietic stem cells (derived from

the mesoderm) may be able to generate both skeletal muscle (also mesoderm derived) and neurons (ectoderm derived) [26].

Mesenchymal Stem Cells

Mesenchymal Stem Cells (MSCs) are multipotent adult stem cells that have the potential to differentiate to lineages of mesenchymal tissues, including *bone* (osteogenic cells), *fat* (adipocytes), *muscle* (myocytes), *cartilage* (chondrocytes), *tendon* (tenocytes) and hematopoiesis-supporting bone-marrow *stroma cells* [27].

Mesenchymal stem cells are mainly derived from the bone marrow stroma (complex array of supporting structures), but they were also isolated from peripheral blood [28], umbilical cord blood [29] and adipose tissues [30].

MSCs were originally isolated from bone marrow aspirate based on their tendency to adhere to a plastic substrate in the cell culture plate, whereas most other bone marrow derived cells (like the highly researched hematopoietic stem cell that also resides in the bone marrow) do not possess this plastic-adherence property [31].

In order to further distinguish mesenchymal stem cells from hematopoietic cells, the cultured cells can be selected against the hematopoietic characteristic markers CD34, CD45 and CD14. In addition, the cell surface marker CD105 (endoglin) and others, are used as positive selection in order to gain MSC enriched cell population. However, there are no currently known MSC-specific cell surface markers that exclusively identify mesenchymal stem cells. Therefore, isolated MSC populations are still not entirely homogenous [32].

Mesenchymal stem cells can be expanded in vitro for many passages, and still retain their multipotential differentiation. Upon induction of differentiation compounds, MSCs differentiate in-vitro to several different mesenchymal lineages such as bone, cartilage, fat, tendon, muscle, and marrow stroma [27].

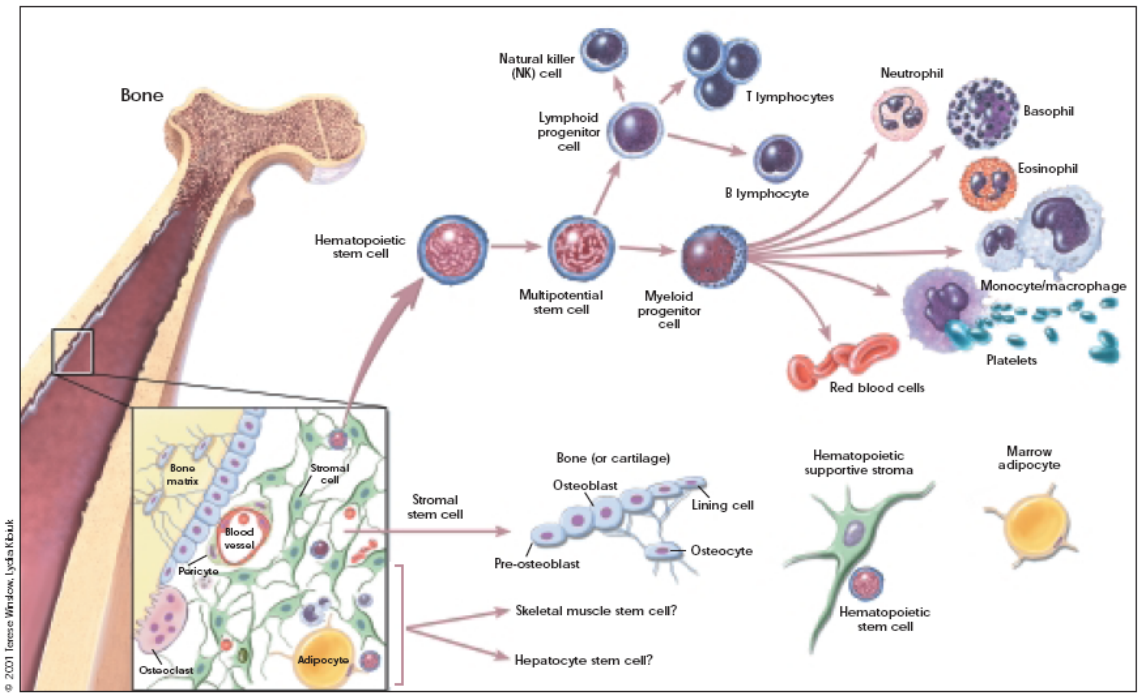


Figure 3. Differentiation of bone marrow derives adult stem cells. Hematopoietic stem cells give rise to the many different types of mature blood cells. Mesenchymal stem cells, derived from the bone marrow stroma can give rise bone, fat and stroma cells.

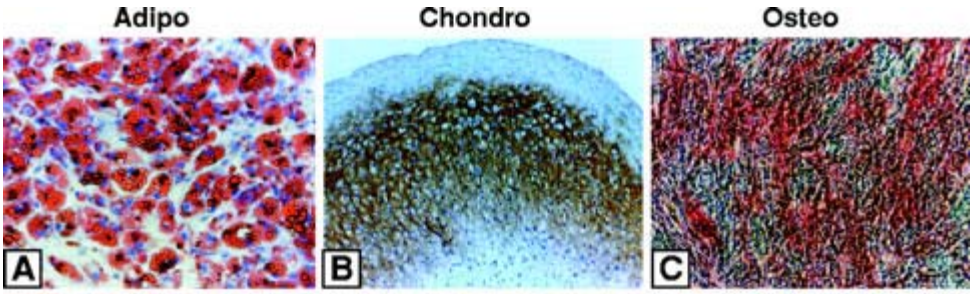


Figure 4. Isolated marrow-derived stem cells differentiate to mesenchymal lineages.
A. Adipocytes (Fat), indicated by the accumulation of neutral lipid vacuoles that stain with oil red; **B**. Chodrocytes (Cartilage), indicated by staining with the C4F6 monoclonal antibody to type II collagen and by morphological changes; **C**. Osteocytes (Bone), indicated by the increase in alkaline phosphatase and calcium deposition.

It was discovered that under certain culturing conditions, mesenchymal stem cells can also "trans-differentiate" to mature specialized cells other than those of the mesenchymal tissues, including neurons, cardiomyocytes and others [26].

This work will focus on mesenchymal stem cells' differentiation into bone and fat mature cells.

Importance of stem cell research

The self-renewal and multipotent differentiation capacity of both embryonic stem cells and adult stem cells make them highly valuable for promoting our understanding of basic developmental processes, and for the development of new revolutionary therapeutic methods.

Stem cells also have potential applications in toxicology and pharmacology, where they can be used to generate mature tissue of different types that may be used for screening of pharmacological compounds [33].

Many diseases (like leukemia) involve a depletion of the stem cell pool in charge of supplying new specialized cells to different mature tissues. Other diseases (like diabetes, Alzheimer and Parkinson) involve destruction or wearing out of tissues as a result of trauma or inadequate replenishment from stem cells pools [33].

Once we know how to control the development of cultured stem cells, we may be able to induce directed differentiation that would yield specific types of mature cells that are required for replacing damaged tissues.

Embryonic stem cells have raised a lot of controversy, as their extraction from young embryos destroys potential human lives and thus raises ethical dilemmas. In recent years, it was shown that adult stem cells are capable of trans-differentiating to yield many types of mature tissues, and thus may be used instead of embryonic stem cells for therapeutic applications. In addition, since adult stem cells have decreased proliferation capacity and tumorigenecity compared to ESCs, they may be also safer for use [34].

Therefore, the emerging field of regenerative medicine is now making use of stem cells in general and mesenchymal stem cells in particular (which are ideal candidates thanks to their proliferative and versatile differentiation potential). For example, mesenchymal stem cells can be used in tissue-engineering strategies where they can be cultured in-vitro to expand their numbers, incorporated into three-dimensional scaffolds to assume required shape, and then transplanted in-vivo to the injured site. Also, MSCs can be used in cell replacement therapy, in which genetic defects can be cured by replacing the mutant host cells with normal allogeneic donor cells [35].

The many versatile applications of stem cell research may explain the tremendous interest that they have triggered in recent years.

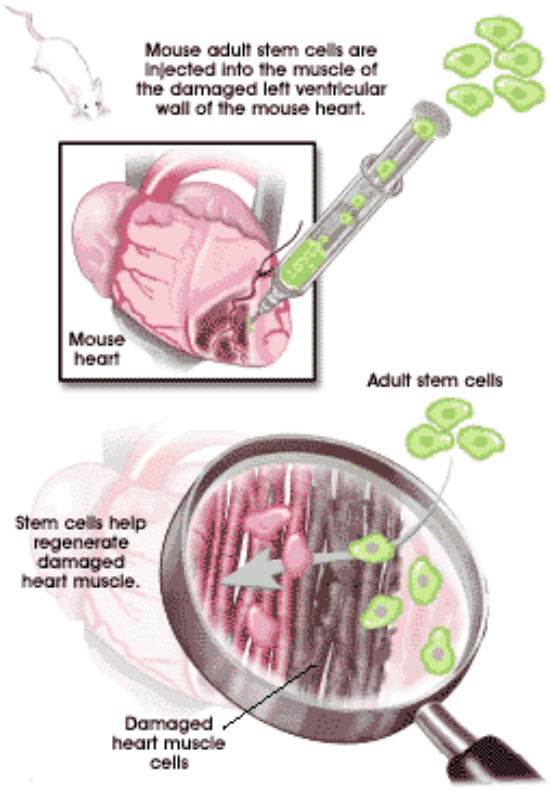


Figure 5. Using adult stem cells to repair damaged heart tissue.

Gene expression and stem cell differentiation

Although most cells in a multi-cellular organism contain the entire genetic information, different cell types express genes in different levels, according to their developmental or functional state. Some genes are found to be highly expressed in most adult tissues ("house keeping" genes), whereas others are highly expressed only on a small subset of adult tissues ("tissue-specific" genes). The subset of expressed genes in a certain time point, determines the properties of the cell and its phenotype.

During differentiation of a stem cell into a mature cell, the cell changes its phenotype as it becomes committed to a certain function by acquiring specific characteristics. A mature osteogenic cell (specialized in aggregating minerals to form the bone) is thus likely to have different transcriptional program compared to a mature adipocytes cell (specialized in fat metabolism). Discovery of genes whose expression is changed along differentiation into a certain lineage may shed light on biological pathways associated with that specific differentiation process and its induction methods.

Little is known about the underlying genetic program that allows stem cells to proliferate for long periods, and yet retain their potential to differentiate into mature cells upon induction. Two attempts to identify "stemness" signature genes common to embryonic, neuronal and hematopoietic murine stem cells have been performed in 2003 [36, 37]. Each study provided a list of genes that allegedly provide stem cells with their unique properties capacities. However, only 6 genes appeared in both gene lists [38, 39]. The small number of common genes may be ascribed to differences in isolation methods, type of computational analysis used to identify shared genes, or differences in microarray chip used in the analysis [23].

On a higher level, several studies tried to detect global expression patterns characterizing embryonic and adult stem cells compared to differentiated cells.

Those studies suggest that stem cells express more genes compared to differentiated cells. Transcription profiling has revealed that most differentiated cell types express only 10–20% of the genome, whereas ESCs express 30-60% of their genes [23].

Golan-Mashiach et al. [40] compared gene expression levels of embryonic, hematopoietic and keratinocyte stems cells with differentiated hematopoietic and keratinocyte tissues, and found a notable down-regulation of genes along the differentiation pathway, accompanied By up-regulation of a smaller set of genes that are needed by the target tissue..

These observations are consistent with the 'priming' hypothesis by which stem cells promiscuously express many different lineage-specific genes at low levels [41]. This transcriptional profile may exist due to the relative open and accessible chromatin state in stem cells compared to mature cells [42]. A transcriptionally permissive chromatin structure may provide stem cells with a rapid differentiation potential when needed during development or in response to an injury [23]. This hypothesis was also named "Just In Case"; stem cells express many genes just in case they are needed in the future (contrary to the parsimonious "just in time" strategy where genes are expressed only when they are needed) [40].

Research Question

- **Global dataset gene expression exploration**
    - What are the prominent differentiation dependant expression patterns observed in the data?
    - How many genes go up/down along the differentiation pathway?
    - Exploration of internal relationships between the samples types.
- **Identification of biological themes, pathways and genes involved in mesenchymal differentiation**
    - What are the major pathways taking part in the differentiation process?
    - What genes play a pivotal role during the differentiation pathway?

Materials and Methods

Embryonic stem cells

Cited from Gerecht-Nir et al., 2003 [43]: "Nondifferentiating hESC lines H9.2 were grown as previously described (Gerecht-Nir et al.,2003 [44]). In brief, the cells were grown on mouse embryonic fibroblasts and passaged every 4 to 6 days using 1 mg/ml type IV collagenase (Gibco Invitrogen Co., San Diego, CA). hESCs were removed from the feeder layers using 1 mg/ml type IV collagenase, further dissociated into small clumps by using 1,000-μl Gilson pipette tips, and cultured in suspension in 50-mm nonadherent Petri dishes (Ein-Shemer, Israel). For analysis, hESC were separated from the feeder layer by type IV collagenase treatment followed by microscopic inspection for the absence of contamination by feeder cells".

Mesenchymal cells

The following section, elaborating mesenchymal stem cell isolation and differentiation, describes the work of Hadi Haslan from Prof. Dan Gazit's Skeletal Biotech lab at the Hebrew University, Jerusalem.

Bone marrow samples were derived from three donors undergoing orthopedic surgery under general anesthesia. Subjects did not suffer any hematological deficiencies. Samples were collected from femur or iliac crest during surgery.

Mesenchymal stem cells were immuno-isolated from bone marrow samples based on the CD105 cell surface marker, after being grown in culture for one week. Then, the mesenchymal stem cells were cultured in different culture media, in order to induce one of the several desirable differentiation states including: no differentiation (minimally cultures MSCs), osteogenic differentiation and adipogenic differentiation.

Donor ID	Age	Gender	Derived samples
#236	63	male	(A) Mesenchymal stem cells
#220	83	female	(B) Bone Control
			(C) Bone Induction
#225	76	male	(D) Fat Control
			(E) Fat Induction

Table 2. Mesenchymal dataset samples

Media used:

2. complete growth medium (**GROW**): DMEM (low glucose) + 10% FCS
3. bone induction medium (**OSTEO IND**): DMEM (low glucose) + 10% FCS + osteogenic supplements
4. bone control medium (**OSTEO CTRL**): DMEM (low glucose) + 10% FCS + buffers used to dissolve the osteogenic supplements
5. fat induction medium (**ADIPO IND**): DMEM (high glucose) + 10% FCS + adipogenic supplements
6. fat control medium (**ADIPO CTRL**): DMEM (high glucose) + 10% FCS + buffers used to dissolve the adipogenic supplements

Culturing details:

1. **Minimally cultured hMSCs – donor#236 – male 63 years old.**

   Day 0: Cells were first plated and grown for 1 week in GROW medium.

   Day 7: Immuno-isolation using CD105 and replating in culture in GROW medium.

Day 17: Isolation of RNA and sending to GeneChip (yielding sample type A).

2. Osteogenic differentiation – donor#220 – female 85 years old.

Day 0: Cells were first plated and grown for 1 week in GROW medium.

Day 7: Immuno-isolation using CD105 and replating in culture in GROW medium until ->

Day 22: Start of osteogenic differentiation: cells replated, addition of OSTEO IND medium or OSTEO CTRL medium.

Grown with the above medium for 2 weeks until ->

Day 36: Isolation of RNA and sending to GeneChip (yielding sample types B and C).

3. Adipogenic differentiation – donor#225 – male 76 years old.

Day 0: Cells were first plated and grown for 1 week in GROW medium.

Day 12: Immuno-isolation using CD105 and replating in culture in GROW medium until ->

Day 26: Start of adipogenic differentiation: cells replated, addition of ADIPO IND medium or ADIPO CTRL medium.

Grown with the above medium for 4 weeks until ->

Day 54: Isolation of RNA and sending to GeneChip (yielding sample types D and E).

Microarrays production

Three Affymetrix Human Genome U133A version 2.0 microarrays were produced for each one of the above mentioned cell types (there are six different samples types: ESCs, MSCs, Bone control, Bone induction, Fat control, Fat Induction). One microarray was damaged (Bone control), leaving 17 microarrays composing the analyzed dataset. The U133Av2 microarray contains 22,215 probesets representing 14,500 well-characterized genes. Affymetrix Microarray Suite Software (MAS, version 5) was used to process the raw microarray data, yielding the scaled data that was used for the bioinformatics analysis.

Dataset Structure Scheme

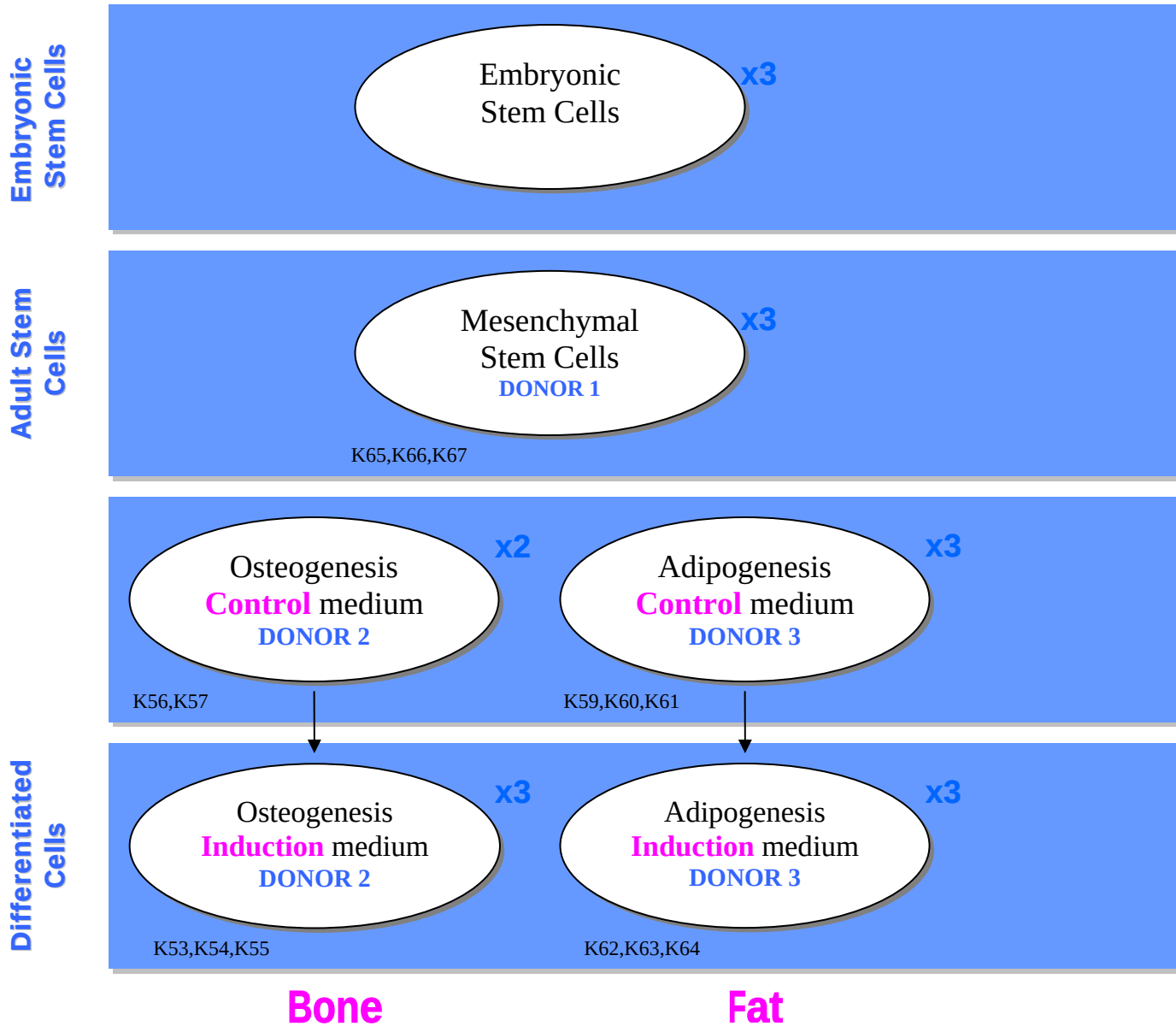


Figure 6. Dataset structure

Results

Global gene expression analysis along the differentiation pathway

We start by examining global expression patterns on the microarray level, trying to determine whether ESC, MSC or differentiated mesenchymal samples differ significantly in the number of genes they express highly.

For this step of the analysis, the data was preprocessed in a permissive manner in order to keep a large number of genes to work with. Starting with 22,215 probe-sets on the original dataset, 5,292 probe-sets were filtered out for being 'absent' on all 17 samples composing the dataset, leaving 16,923 probe-sets. A threshold of 1 followed by log2 transformation was then applied to the data.

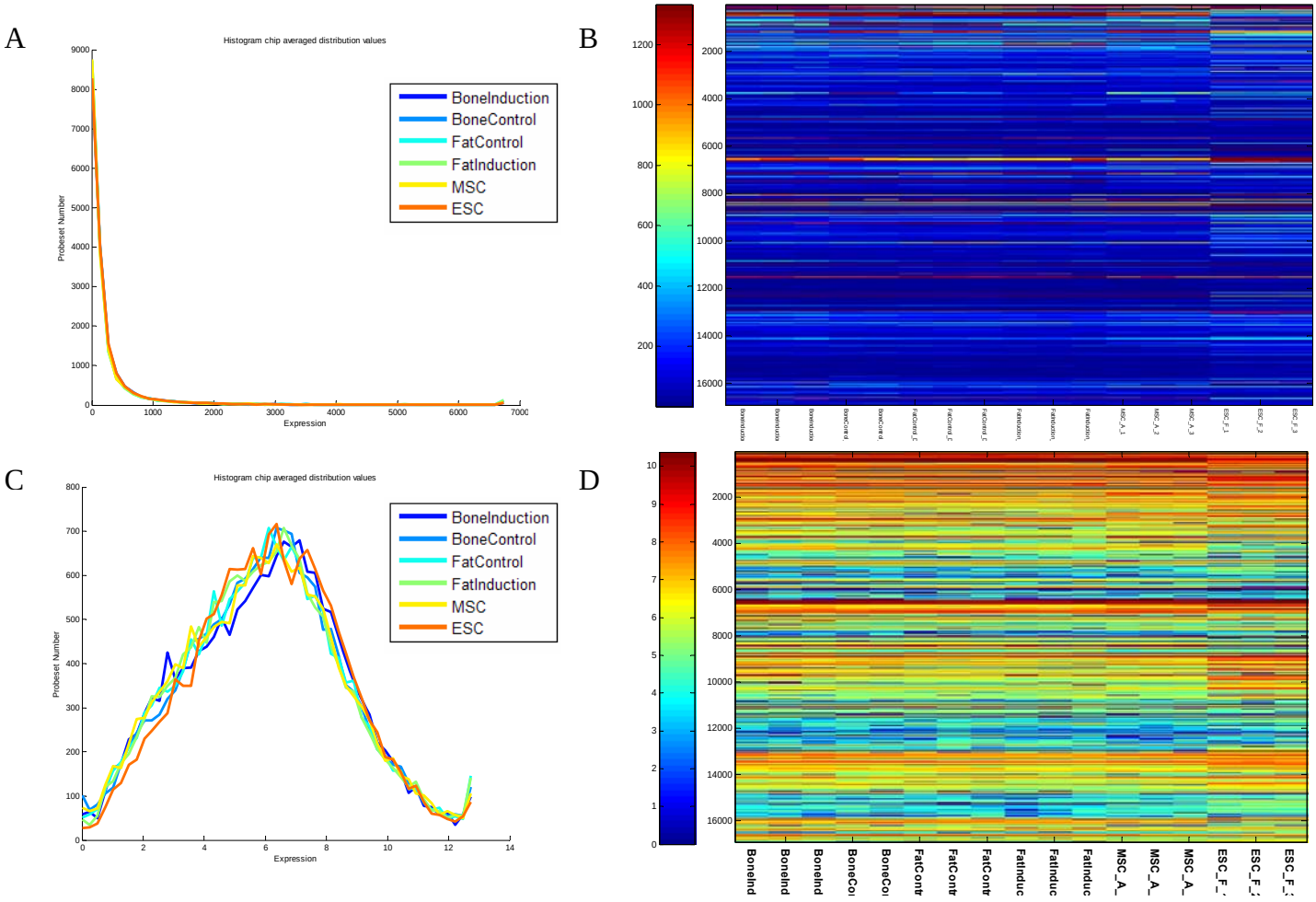


Figure 7. Dataset preprocessing. Plots A and C on the left show expression value histograms for the 6 sample groups (replicates are averaged) before (A) and after (C) applying a threshold of 1 and log2 transformation. The images on the right show the corresponding expression matrices before (B) and after (D) applying the threshold and log2. In the expression matrices, red represents high values, blue represents low values. The color bar shows the mapping of values to colors.

A brief examination of the expression matrix reveals that the three rightmost columns, representing the embryonic stem cell samples, are quite different from the other 14 mesenchymal samples.

In order to examine gene expression similarity between pairs of sample types (e.g. ESCs versus MSCs), we have first averaged the replicates of each sample type, and then subtracted the first average from the second average for each gene. Since the data is log2-transformed, the difference values we got are equivalent to fold change of each gene between the two sample types.

The following two histograms and accompanying tables (the first focuses on fat samples and the second on bone) summarize the distribution of the differences we have calculated. In general, narrow distributions (low standard deviation) are typical of many very small differences and thus indicate high microarray-level similarity between the pair samples. Wide distributions indicate that many genes have a high fold change between the two sample types and are therefore correlated with significant expression level differences.

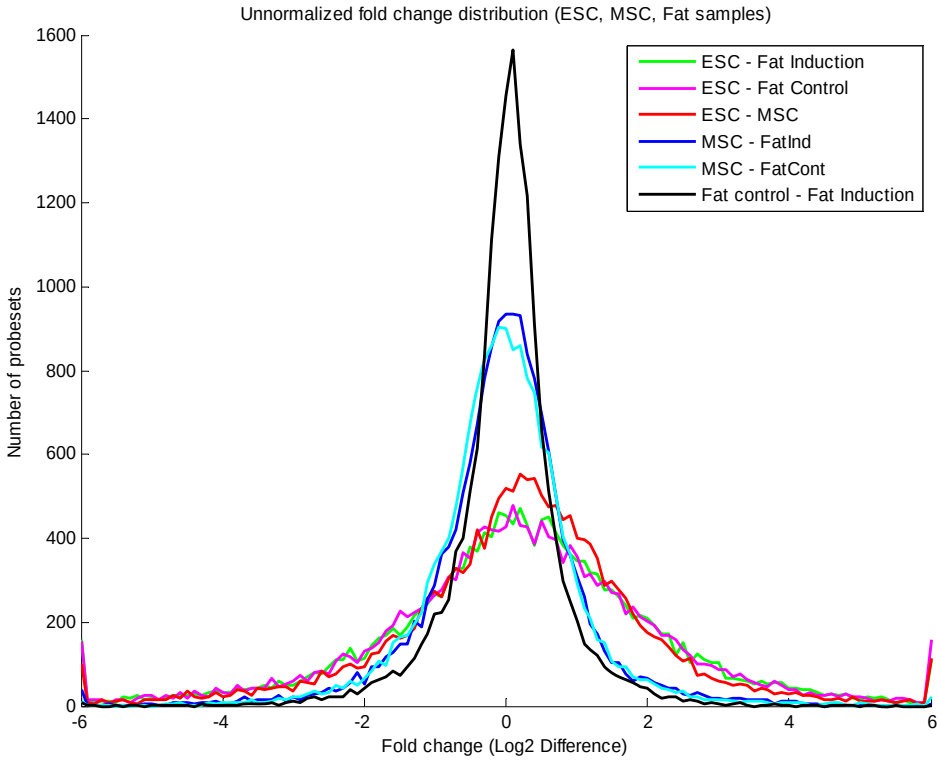


Figure 8. Distribution of probe-set expression difference between sample pairs including ESC, MSC and Fat samples.

Compared pair	Diff. Mean	Diff. STD	Rank sum pValue
ESC – Fat-Induction	0.185	2.070	1.75e-013
ESC – Fat-Control	0.182	2.093	5.90e-012
ESC - MSC	0.216	1.862	7.10e-015
MSC – Fat-Induction	-0.031	1.151	0.57
MSC- Fat-Control	-0.033	1.092	0.34
Fat-Control – Fat-Induction	0.002	0.809	0.68

Table 3: Sample comparison. Mean and standard deviation relate to the difference distribution of the corresponding sample pair, as plotted above. p-value refers to Wilcoxson rank sum test conducted on replicate-averaged expression values of first sample type versus the second (test conducted on expression values, not on difference).

Analysis of these histograms and tables reveals the following:

- **Fat-control and Fat-Induction samples are highly similar on the microarray-level** (The black histogram exhibits the lowest standard deviation). This is expected, as both samples come from the same donor, and their culturing media are very similar (differ only in the presence induction compounds).
- **Embryonic stem cells differ significantly from the three other cell types** (the red, magenta and green histograms exhibit the highest variances). This suggests that the expression of many genes is either lower or higher on the ESC samples compared to MSC, Fat-Control and Fat-induction samples. ESCs therefore express many genes in an *extreme* (either lower or higher) manner compared to MSC, Fat-Control and Fat-Induction samples. Since the areas under the red, magenta and green curves are larger for positive difference values compared to negative values, we concluded that the number of genes that are expressed by ESCs at a higher level than by the mesenchymal cells exceeds significantly the number of genes with the opposite difference pattern.
- **Mesenchymal stem cells exhibit medium microarray-level similarity to the Fat-Induction and Fat-Control samples**. Dark blue and light blue curves represent comparison of mesenchymal stem cells

with Fat-Induction and with Fat-Control respectively. These two difference histograms display a distribution similar to the black curve representing the Fat-Control versus Fat-Induction comparison, but their standard deviation is larger. This suggest that MSCs somewhat differ on the microarray-level from both Fat-control and Fat-Induction; however MSCs are still more similar to the two fat sample types compared to ESCs.

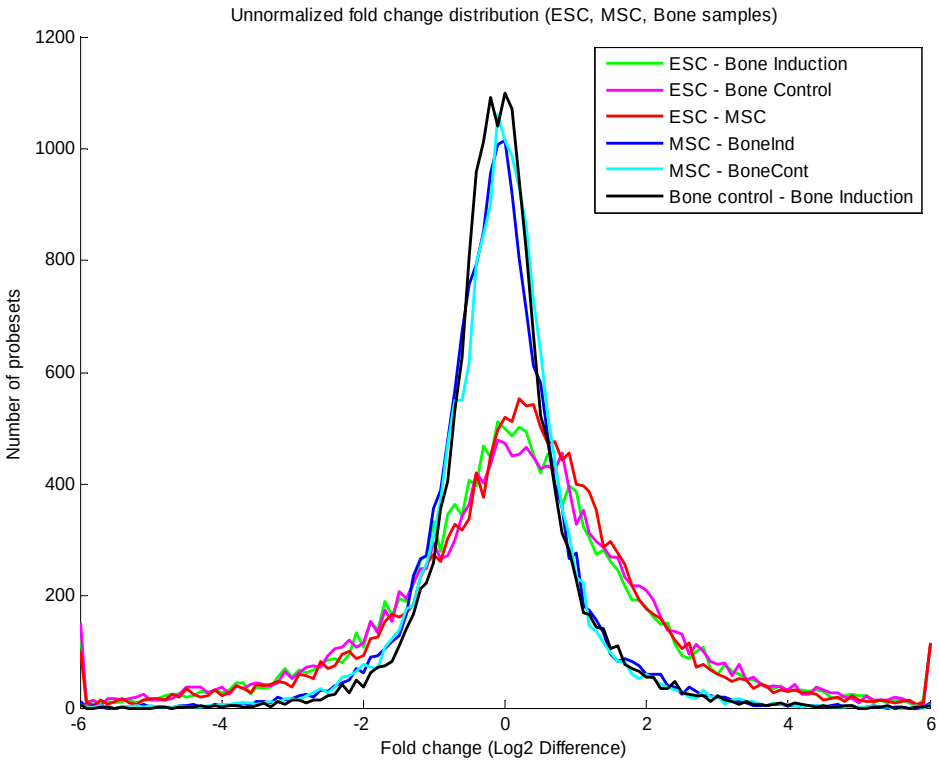


Figure 9. Distribution of probe-set expression difference between sample pairs, including ESC. MSC and Bone samples.

Compared pair	Diff. Mean	Diff. STD	Ranksum pValue
ESC – Bone-Induction	0.136	1.947	3.08e-005
ESC – Bone-Control	0.157	2.022	2.64e-008
ESC - MSC	0.216	1.862	7.05e-015
MSC – Bone-Induction	-0.080	1.046	0.0005
MSC- Bone-Control	-0.058	1.004	0.02
Bone-Control – Bone-Induction	0.002	0.809	0.17

Table 4. Sample comparison. Mean and standard-deviation relate to the difference distribution of the corresponding sample pair, as plotted above. p-value refers to Wilcoxson rank sum test conducted on replicate-averaged expression values of first sample type versus the second (test conducted on expression values, not on difference).

Similarly, examination of the above histogram and table raised the following observation:

- **Bone-Control and Bone-Induction samples have the highest level of similarity** (the black histogram exhibits the smallest standard deviation).
- **ESC samples compared with MSCs, Bone-Control and Bone-Induction show large dissimilarity** (red, magenta and green curves). ESCs express many genes at lower levels than in the other samples. ESCs also express an even larger number of genes at higher levels than in the other samples.
- **MSCs exhibit high similarity to the Bone-control and Bone-induction samples, closer than MSCs are to Fat-Control and Fat-Induction**.

The rank-sum p-value columns in the last two tables reveal that the mean expression (after averaging replicates) of ESCs differs significantly from the mean of all five mesenchymal sample types, using a significance level of 0.05.

Interestingly, the only mesenchymal sample type that differs significantly from mesenchymal stem cells is the Bone-Induction sample type. This observation may be explained by higher biological difference induced during osteogenesis, or by basal genetic differences between the two different donors whose tissues were used to prepare the MSC and Bone samples.

When applying standardization on the dataset rows (centering and normalizing the 16,923 probe-sets), the global dissimilarity between the embryonic stem cell samples and other mesenchymal samples is prominently demonstrated. The standardization changes the distribution of the embryonic stem cell samples to a bi-modal distribution, emphasizing that ESCs express many genes at a lower level than the mesenchymal samples, and even more genes at a higher level than the mesenchymal samples.

This remarkable change in ESC expression value distribution due to standardization may be explained by the existence of many rather small differences in expression between the ESC samples and the other mesenchymal samples that are being magnified by the standardization transformation. For more on this issue, please refer to Appendix I.

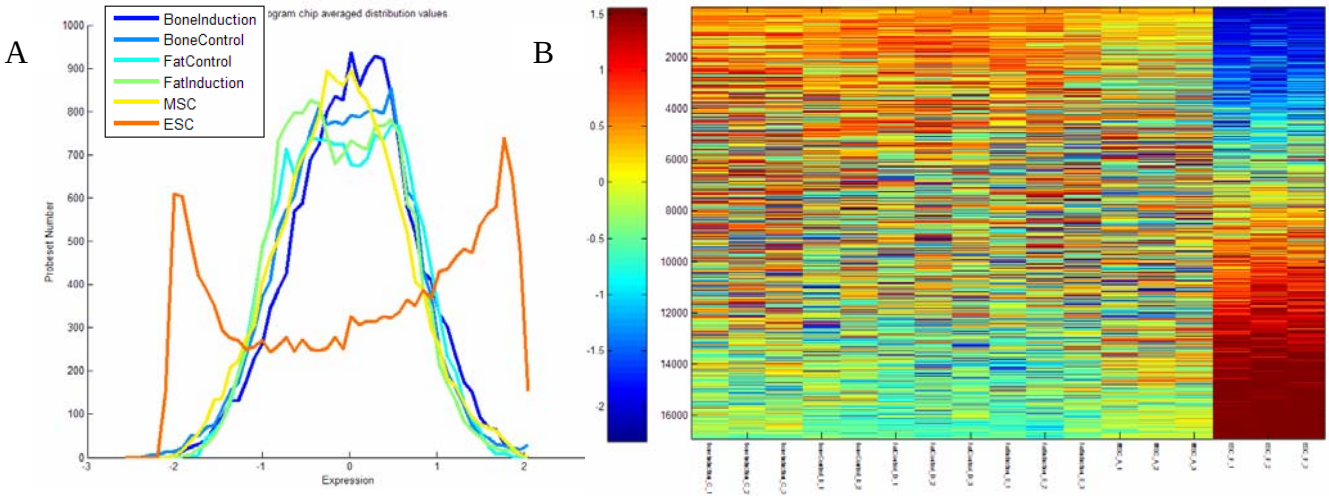


Figure 10. The effect of standardization on sample distribution. (A) Standardized expression distribution for the 6 sample types (replicates are averaged). (B) The corresponding expression matrix. The last three columns represent the Embryonic stem cells. Matrix genes are sorted according to ESC replicates average.

Counting differentially expressed genes

In order to asses the number of differentially expressed genes significantly varying between the dataset samples, we have conducted a statistical test aimed at filtering out unvarying genes and then compared gene expression between the sample types.

For this analysis, a more stringent set of preprocessing parameters was used. As before, 16,962 probe-sets were left after removing 5,292 'all-absent' probe-sets from the initial 22,215 probe-set dataset. A threshold of 16 was applied to any expression values lower than this threshold, and the data was log2-transformed. 586 probe-sets were detected as having a standard deviation of zero, and were removed. The remaining 16,337 probe-sets were used for this analysis, and for several of the subsequent analysis steps.

We have used one-way ANOVA (Analysis of Variance) to keep only genes that vary significantly between sample groups (their variance between groups is larger than the total variance within the groups). FDR of 0.05 was then applied on the six sample groups, yielding 12,461 differentially expressed probe-sets.

In the following figure, ESC and MSC expression values for the 12,461 differentially expressed probe-sets were plotted (replicates were averaged), sorted by their expression on the ESC samples. The black line is formed by the many dots representing the sorted ESC probesets. For each black dot, there is a vertically corresponding blue dot, representing the probe-set's averaged expression on the MSC samples.

Counting the blue dots above, below and on the black line yielded the following:

6980 probe-sets are higher on ESC compared to MSC (dots under the line)

5302 probe-sets are lower on ESC compared to MSC (dots above the line)

179 probe-sets are equal on both ESC and MSC (dots on the line)

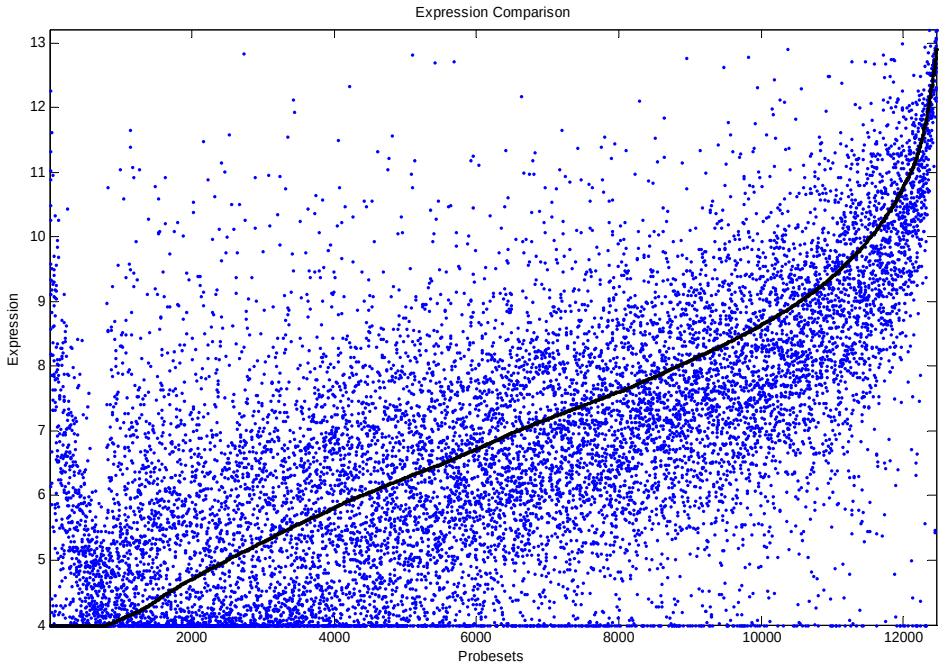


Figure 11: Comparing expression levels of ESC and MSC differentially expressed genes. 12,461 differentially expressed probe-sets that passed the ANOVA test over the six sample types, with FDR of 0.05, are plotted along the X-axis, sorted by their averaged expression on the ESC samples. Black dots represent probe-set expression on the ESC samples, blue dots represent expression on the MSC samples.

In the above plot, there are 1,678 more blue dots under the black line compared to blue dots above the line, indicating that out of the 12,461 differentially expressed genes, 1,678 (13.4%) are expressed higher on ESC compared to MSC.

In a similar manner, we have compared MSC averaged expression on the 12,461 differentially expressed genes to the averaged expression of the Fat-Induction and Bone-Induction samples as can be seen on the next figure (Black – MSC; Red – Fat-Induction; Blue – Bone-Induction).

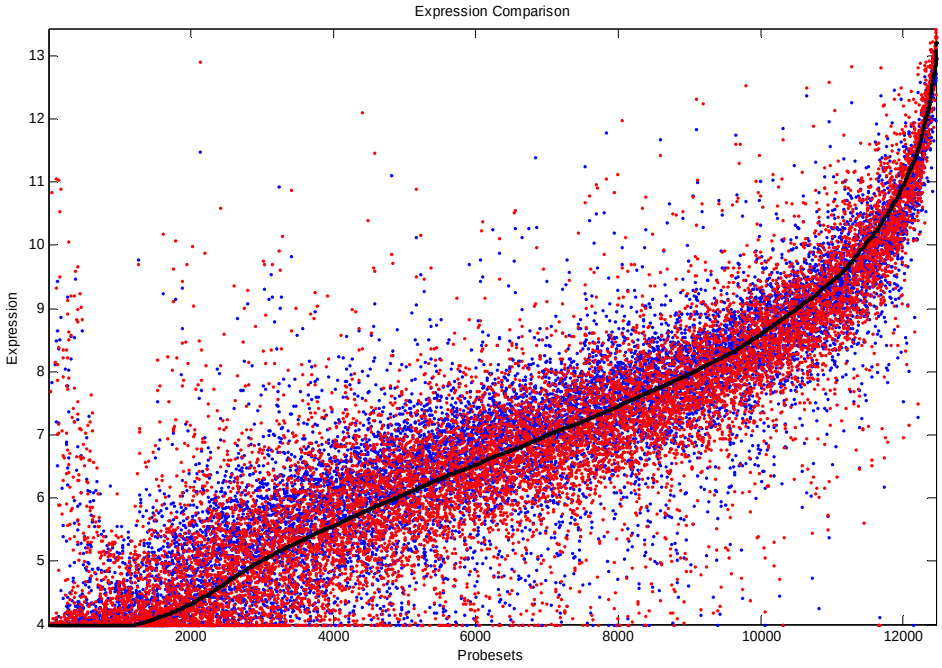


Figure 12: Comparing expression levels of MSC with Fat-Induction and Bone-Induction differentially expressed genes. 12,461 differentially expressed probe-sets that passed the ANOVA test over the six sample types, with FDR of 0.05, are plotted along the X-axis, sorted by their averaged expression on the MSC samples. Black dots represent probe-set expression on the MSC samples, red dots represent expression on the Fat-Induction samples and blue dots represent expression on the MSC samples.

Counting the blue dots in the above figure reveals that 5160 probe-sets are higher, 765 are equal and 6536 are lower on MSC compared to Bone Induction.

Counting the red dots in the above figure reveals that 6248 probe-sets are higher, 682 are equal and 5531 are lower on MSC compared to Fat Induction.

In total, out of 12,461 differentially expressed probe-sets, Fat-Induction samples have 717 (5.7%) more over-expressed probe-sets than under-expressed probe-sets compared with MSC samples. On the other hand, Bone-Induction samples have 1376 (11%) more under-expressed probe-sets than over-expressed probe-sets compared with MSC samples.

The following summary table shows the net difference between over-expressed and under-expressed probe-sets, calculated for more sample pairs. The entries with yellow background are those already brought above in detail.

	MSC	FatControl	FatInduction	BoneControl	BoneInduction
ESC	1678	1286	1472	1331	916
MSC		27	717	-448	-1376
FatControl			1128	-270	-1342
BoneControl		270	863		-1411

Table 5: Summary of expression difference between pairs of sample types. Each entry represent the number of over-expressed probe-sets subtracted by under-expressed probe-sets in the sample type specified in the column header, compared to the sample type specified in the row header.

To summarize this part, the various visualizations and calculations enabled us to determine the following:

- ESCs express more over-expressed probe-sets compared with all mesenchymal sample types.
- MSCs express more over-expressed probe-sets compared with Fat samples, but more under-expressed probe-sets compared with Bone samples.

Identification of genes changed upon induction

Differentiation of MSC to Fat

Zooming into the differentiation process by which mesenchymal stem cells develop into differentiated fat cells, we are interested in identifying genes that significantly change upon induction to fat. Since the MSC samples were taken from one donor, and both Fat-Control and Fat-Induction samples were taken from a second donor, comparing MSC samples directly with Fat-Induction samples may detect genes that differ due to differentiation induction, but will probably also yield genes that differ between the two sample types due to *donor variance*. Considering this, we have decided to look for genes that significantly change between Fat-Control and Fat-Induction (taken from the same donor), but that also exhibit no change, or a corresponding change between MSC and Fat-Control. To pinpoint such genes we have first used t-test to find genes whose expression mean significantly vary between Fat-Control to Fat-Induction samples, and then used *profile filtering* to remove genes based on their fold change between MSC to Fat-Control.

To this end, we have used the MSC, Fat-Control and Fat-Induction samples from the permissively preprocessed dataset of the first analysis above (preprocessing includes removal of 'all-absent' probe-sets, threshold of 1 and log2 transformation). T-test was independently applied on each one of the 16,932 probe-sets, testing for equality of means between Fat-Control (3 samples) and Fat-Induction (3 samples). FDR of 0.05 was then applied on the t-test p-values, yielding a list of 651 probe-sets, whose mean significantly differs between the Fat-Control samples and Fat-Induction samples.

We then applied the *expression profiling* method using a resolution of N=10 on the 651 differentially expressed probe-sets. Expression profiles were calculated by mapping the replicate-averaged expression of each probe-set to one of 10 bins. Probe-set expression was then represented by a three component vector (MSC, Fat-Control, Fat-Induction) using 10 levels of expression for each

component. Bin intervals were defined by dividing dataset expression range (2nd percentile - 98th percentile) into 10 intervals. In this case, difference between two sequent expression bins is equivalent to a fold change of ~2 (2^1.07) as (11.7 [98th percentile value] – 1.0 [2nd percentile value] / 10 [bins] = 1.07. 260 different profiles were identified among the 651 differentially expressed genes.

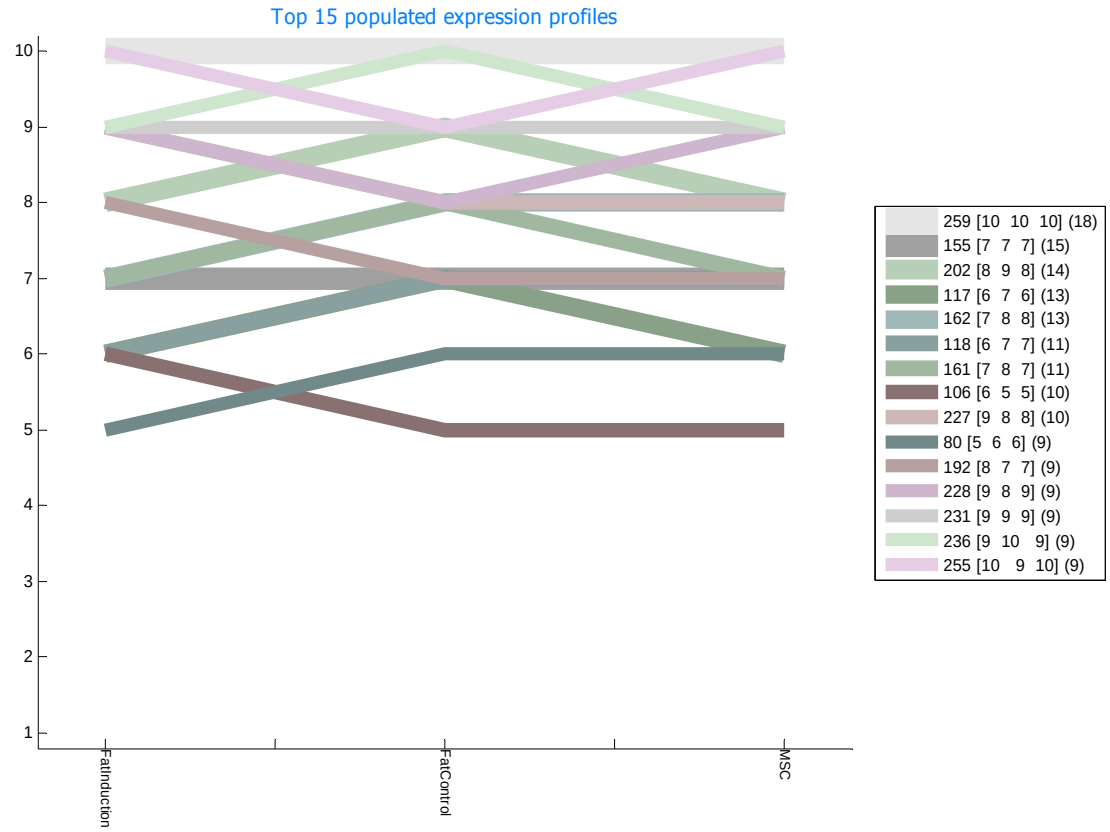


Figure 13. Top 15 populated expression profiles for the MSC, Fat-Control and Fat-Induction samples. Each color ribbon represents a cluster of genes sharing the same expression profile (ribbon's width represent the cluster size). The Y axis spans through the 10 expression bins (profiling resolution is a user-defined parameter). For example, the light-gray ribbon on top (ribbon #259), represent the largest gene cluster. The 18 genes included in this cluster exhibit constant expression as they are mapped to the profile [10 10 10].

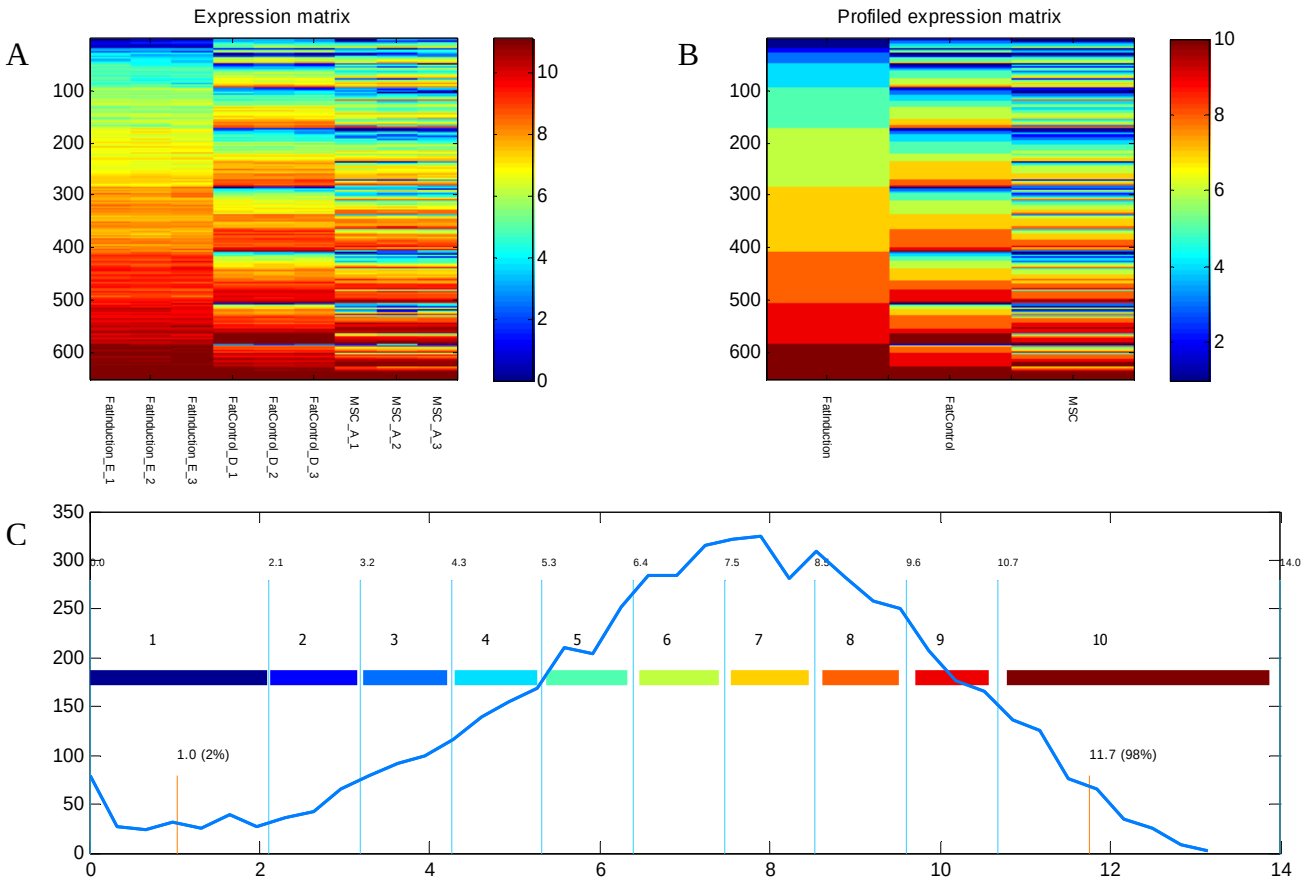


Figure 14. Expression profiling of the MSC, Fat-Control and Fat-Induction samples. (A) Reordered expression matrix based on sorted profiles. (B) Sorted expression profiles. (C) Distribution of gene expression values for the three sample types. Orange vertical lines represent the 2nd and 98th percentiles; this expression range is sliced to 10 bins defining the mapping of expression value to profiles of resolution of 10.

In order to identify differentially expressed genes whose expression is **up-regulated** upon induction to fat, we have filtered in genes whose profile difference series is (>0 >=0), which means keeping only profiles that are higher on Fat-Induction compared to Fat-Control, and that are equal or higher on Fat-Control compared to Mesenchymal stem cells (we have allowed MSC-FatControl difference to be equal or higher than 0 in order to keep enough genes for the subsequence analysis). 88 such profiles (representing 210 probesets) passed the above criteria and they are displayed in figure 15.

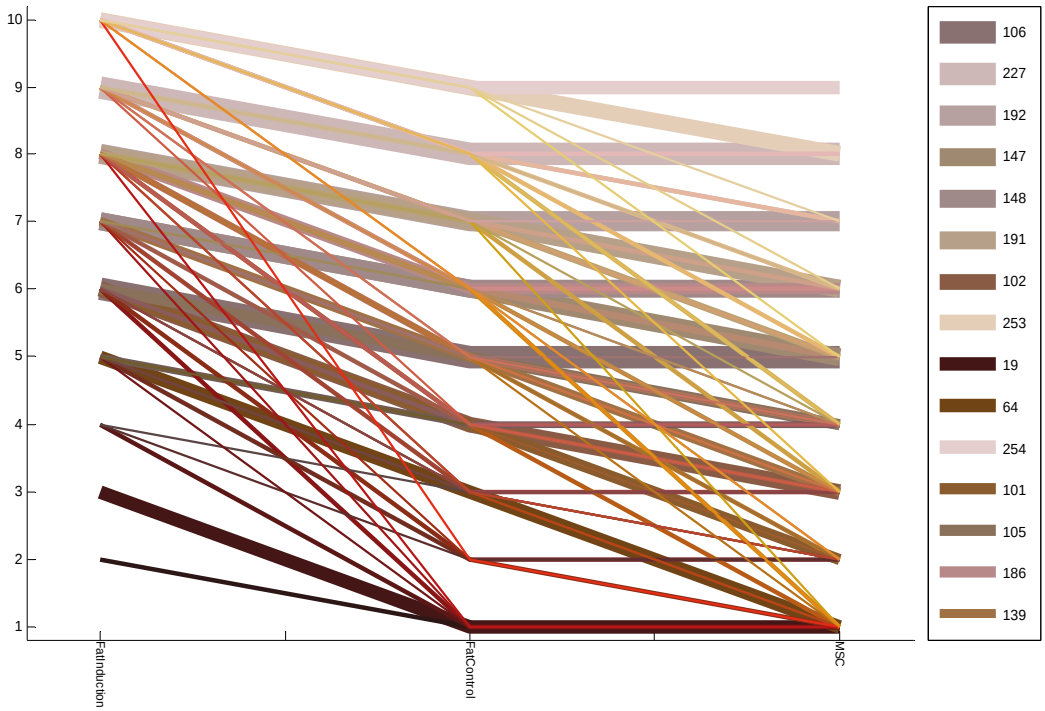


Figure 15. The 88 filtered profiles that are increasing from MSC to Fat-Induction. These profiles encapsulates 210 genes that are up-regulated upon induction to fat. Since the profiling operation was applied on non-normalized data, several profiles may have the same relative expression pattern. Line widths represent the number of probe-sets in the profile.

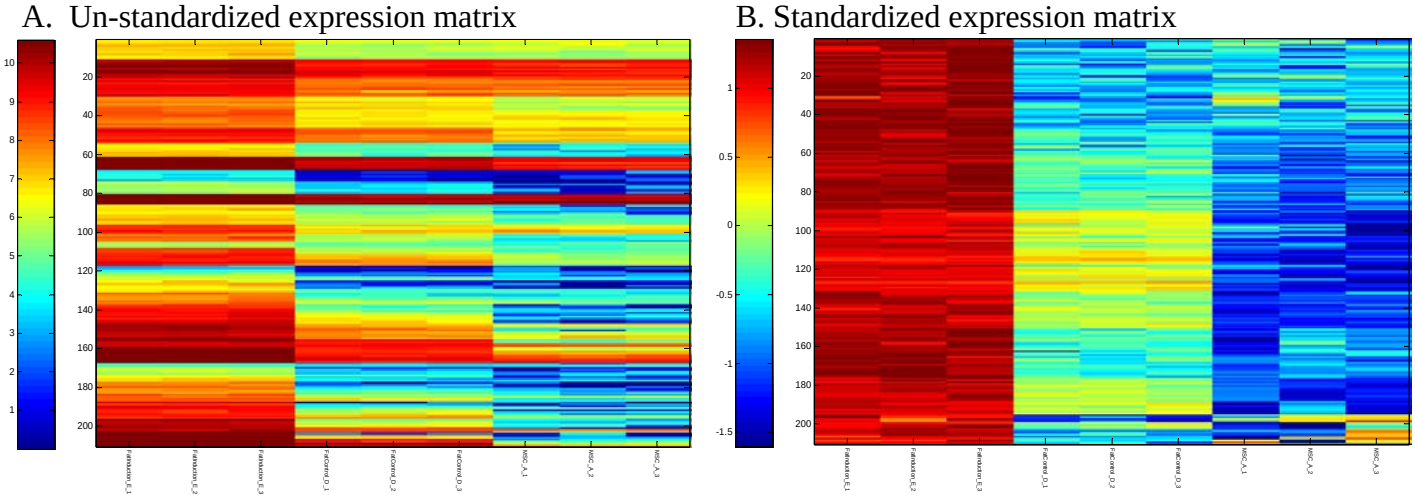


Figure 16. Expression matrices of the 210 probesets whose expression profile is increasing from MSC to Fat-Induction. (A) Un-standardized expression matrix, probe-sets are ordered according to the 88 profiles. (B). Standardized expression matrix, probe-sets are ordered by SPC clustering.

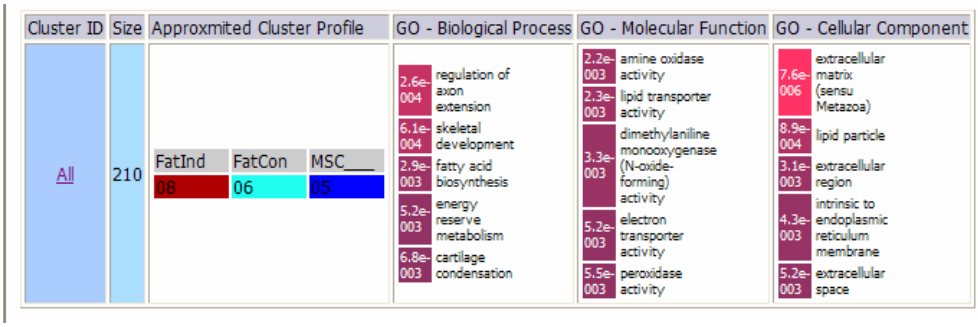


Table 6. Gene ontology enrichment test results. Top 5 significant gene ontology terms the 210 probesets over-expressed upon induction to fat.

Gene ontology enrichment was calculated for the set of 210 probe-sets; complete results (enriched GO terms and full gene list) are published online at http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_up/1_results.html.

After standardization, the 210 probe-sets were clustered using the SPC algorithm. Produced clusters were also tested for gene ontology enrichment. GO enrichment test results and complete gene list are published online at http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_up/1_c_summary.html.

Prominent significant GO terms and included genes:

- *Biological process – fatty acid biosynthesis*

Gene Symbol	Gene Title
SCD	stearoyl-CoA desaturase (delta-9-desaturase)
PTGS1	prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase)
THEDC1	thioesterase domain containing 1
ACACB	acetyl-Coenzyme A carboxylase beta

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_up/20060504T014501_GO_PS_6633.html

- *Biological process – skeletal development*

Gene Symbol	Gene Title
IGF2	insulin-like growth factor 2 (somatomedin A)
FRZB	frizzled-related protein
PRELP	proline/arginine-rich end leucine-rich repeat protein
PTHR1	parathyroid hormone receptor 1
DLX5	distal-less homeo box 5

ALPL	alkaline phosphatase, liver/bone/kidney
NPR3	natriuretic peptide receptor C/guanylate cyclase C (atrionatriuretic peptide receptor C)

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_up/20060504T015048_GO_PS_1501.html

- Molecular function – lipid transporter activity

Gene Symbol	Gene Title
APOD	apolipoprotein D
APOE	apolipoprotein E
SAA1	serum amyloid A1
APOL1	apolipoprotein L, 1
SAA1	serum amyloid A1

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_up/20060504T015202_GO_PS_5319.html

- Cellular component – extra-cellular matrix

Gene Symbol	Gene Title
SPARCL1	SPARC-like 1 (mast9, hevin)
NID1	nidogen 1
MGP	matrix Gla protein
PRELP	proline/arginine-rich end leucine-rich repeat protein
COL11A1	collagen, type XI, alpha 1
MMP19	matrix metallopeptidase 19
LAMA2	laminin, alpha 2 (merosin, congenital muscular dystrophy)
OMD	osteomodulin
CILP	cartilage intermediate layer protein, nucleotide pyrophosphohydrolase
TIMP4	TIMP metallopeptidase inhibitor 4
DPT	dermatopontin
SPON1	spondin 1, extracellular matrix protein
DPT	dermatopontin
MFAP5	microfibrillar associated protein 5
ADAMTS2	ADAM metallopeptidase with thrombospondin type 1 motif, 2

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_up/20060504T015434_GO_PS_5578.html

Turning now to identifying genes that are **down-regulated** upon induction to fat, we have filtered in profiles whose expression Fat-Control is equal or lower compared to MSC, and lower on Fat-Induction compared to Fat-Control. These criteria yielded 37 profiles, representing 97 probe-sets, which we have used for gene ontology enrichment test.

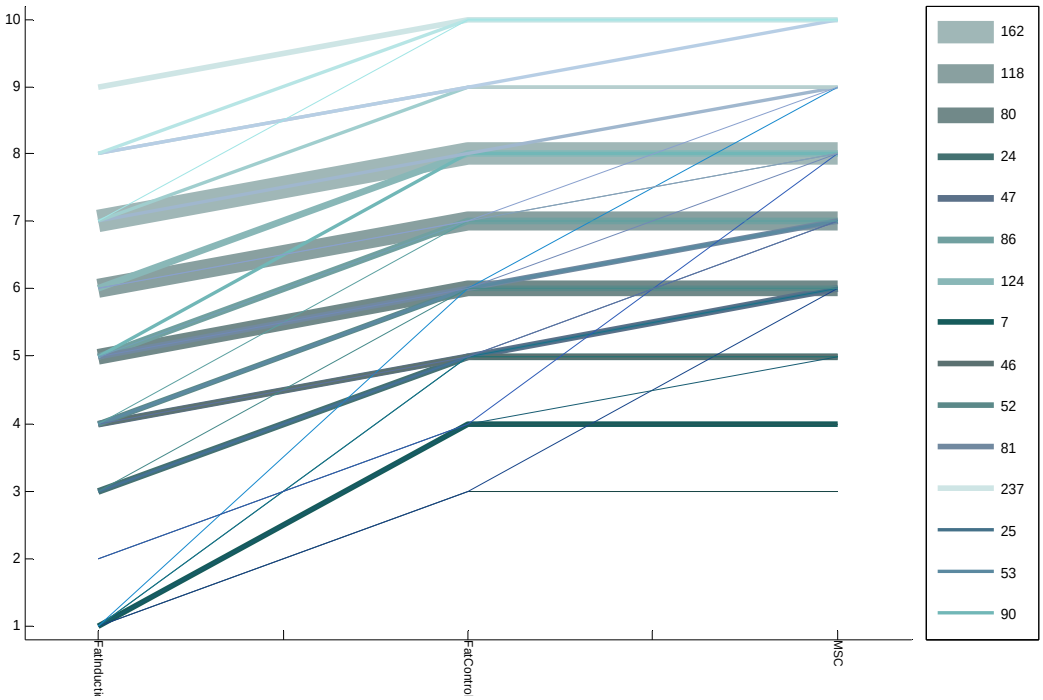


Figure 17. The 37 filtered profiles that are decreading from MSC to Fat-Induction. These profiles encapsulates 97 genes that are down-regulated upon induction to fat.

A. Un-standardized expression matrix

B. Standardized expression matrix

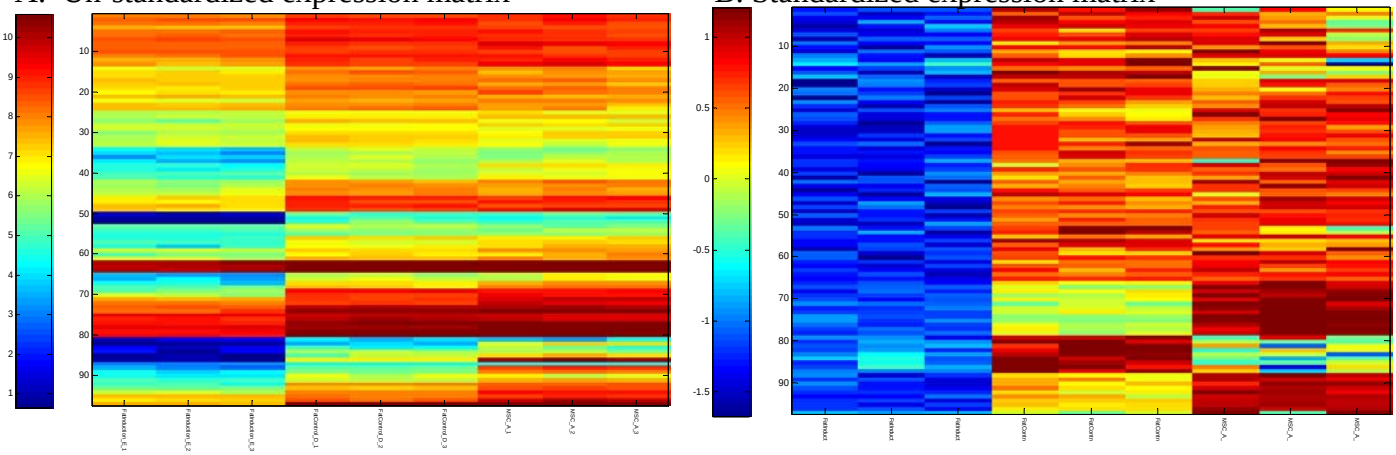


Figure 18. Expression matrices of the 97 probesets whose expression profile is decreasing from MSC to Fat-Induction. (A) Un-standardized expression matrix, probe-sets are ordered according to the 37 profiles. (B). Standardized expression matrix, probe-sets are ordered by SPC clustering.

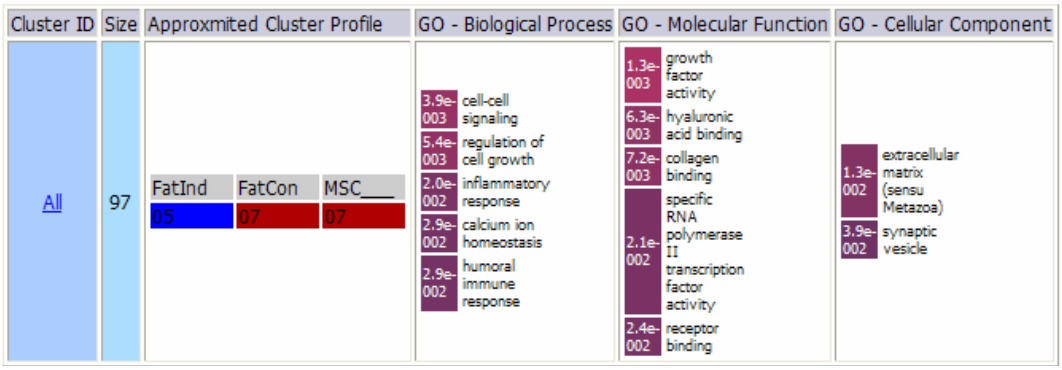


Table 7. Gene ontology enrichment test results. Top significant gene ontology terms for the 97 probesets under-expressed upon induction to fat.

Gene ontology enrichment was calculated for the set of 97 probesets; complete results (enriched GO terms and full gene list) are published online at http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_down/2_results.html.

After standardization, the 97 probe-sets were clustered using the SPC algorithm. Produced clusters were also tested for gene ontology enrichment. GO enrichment test results and complete gene list are published online at http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_down/2_c_summary.html.

Prominent significant GO terms and included genes:

- *Biological process – cell-cell signaling*

Gene Symbol	Gene Title
GJB1	gap junction protein, beta 1, 32kDa (connexin 32, Charcot-Marie-Tooth neuropathy, X-linked)
IL6	interleukin 6 (interferon, beta 2)
ADORA1	adenosine A1 receptor
TNFAIP6	tumor necrosis factor, alpha-induced protein 6
FADS1	fatty acid desaturase 1
MDK	midkine (neurite growth-promoting factor 2)
CXCL12	chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
CCL2	chemokine (C-C motif) ligand 2

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_down/20060504T022500_GO_PS_7267.html

- *Biological process – regulation of cell growth*

Gene Symbol	Gene Title
QSCN6	quiescin Q6
NET1	neuroepithelial cell transforming gene 1

BRD8	bromodomain containing 8
IGFBP3	insulin-like growth factor binding protein 3

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_down/20060504T023004_GO_PS_1558.html

- Molecular function – growth factor activity

Gene Symbol	Gene Title
IL6	interleukin 6 (interferon, beta 2)
NRG1	neuregulin 1
MDK	midkine (neurite growth-promoting factor 2)
CXCL12	chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
HGF	hepatocyte growth factor (hepapoietin A; scatter factor)
PDGFC	platelet derived growth factor C

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/fat_down/20060504T023413_GO_PS_8083.html

Differentiation of MSC to Bone

Similarly to the analysis performed on the fat samples, we have looked for genes whose expression changes upon induction to bone by applying t-test (FDR of 0.05) on the 16,932 dataset probe-sets, testing mean equality between the Bone-Control samples and the Bone-Induction samples; 734 probe-sets passed this test. We have then profiled the 734 differentially expressed probe-sets using the MSC, Bone-Control and Bone-Induction samples into 203 different profiles using a resolution of 10. Figure 19 displays the 15 largest profiles; note that several profiles have the same expression level on the MSC and Bone-Induction samples, but differ in their expression on Bone-Control (profiles #129, #158, #104) – this is exactly the pattern which we are trying to compensate for (considering donor-variance) by applying the mentioned filter on the profiles.

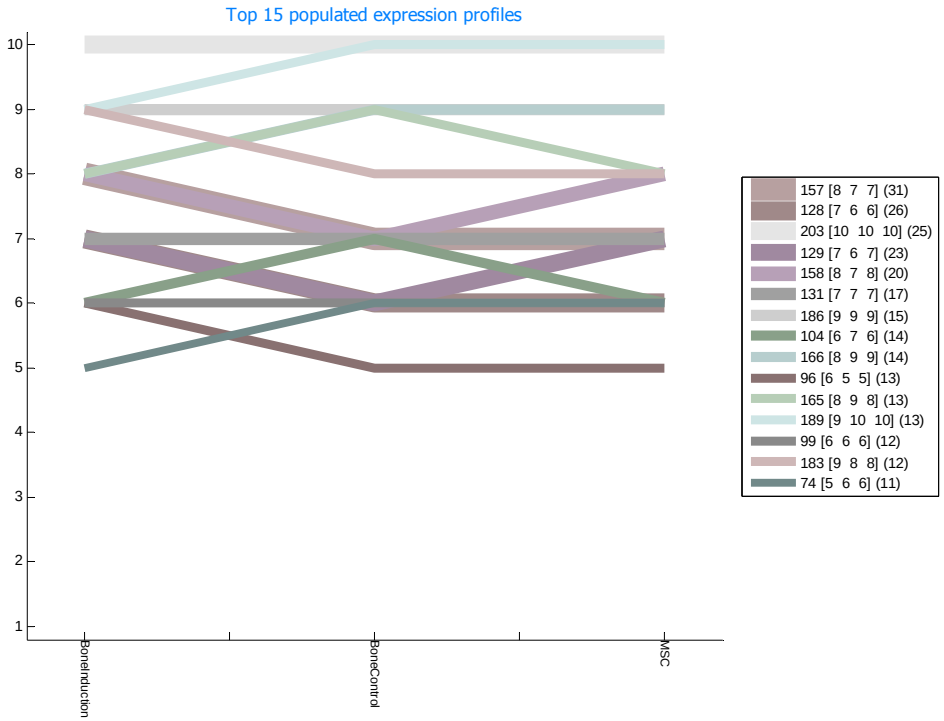


Figure 19. Top 15 populated expression profiles for the MSC, Bone-Control and Bone-Induction samples. Each color ribbon represents a cluster of genes sharing the same expression profile (ribbon's width represent the cluster size). The Y axis spans through the 10 expression bins (profiling resolution is a user-defined parameter).

Starting with genes whose expression rises upon induction to bone, 49 profiles (representing 195 probe-sets) exhibited higher expression on Bone-Induction compared with Bone-Control, and exhibited higher or equal expression on Bone-Control compared with MSC.

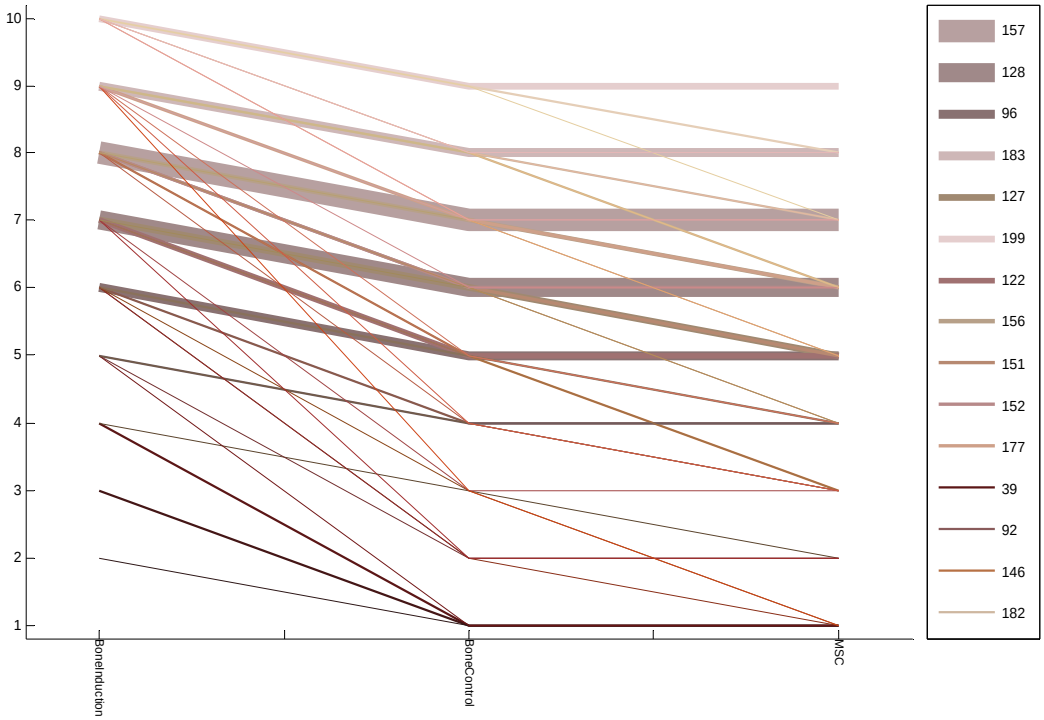


Figure 20. The 49 filtered profiles that are increasing from MSC to Bone-Induction. These profiles encapsulates 195 genes that are up-regulated upon induction to fat. Since the profiling operation was applied on non-normalized data, several profiles may have the same relative expression pattern. Line widths represent the number of probe-sets in the profile.

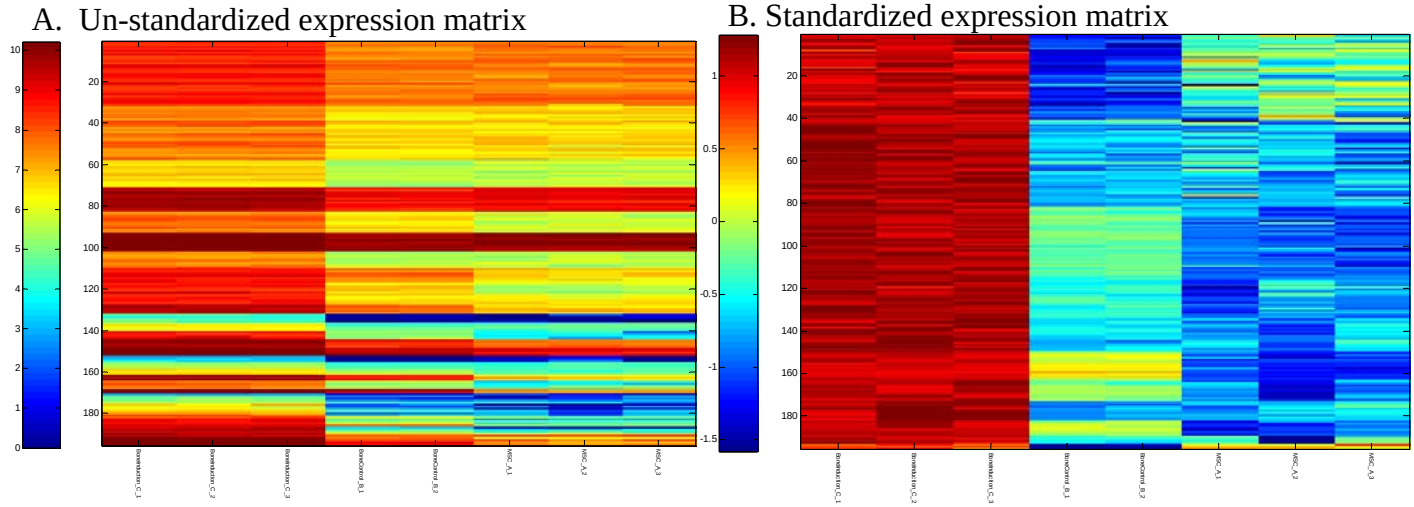


Figure 21. Expression matrices of the 195 probesets whose expression profile is increasing from MSC to Bone-Induction. (A) Un-standardized expression matrix, probe-sets are ordered according to the 49 profiles. (B). Standardized expression matrix, probe-sets are ordered by SPC clustering.

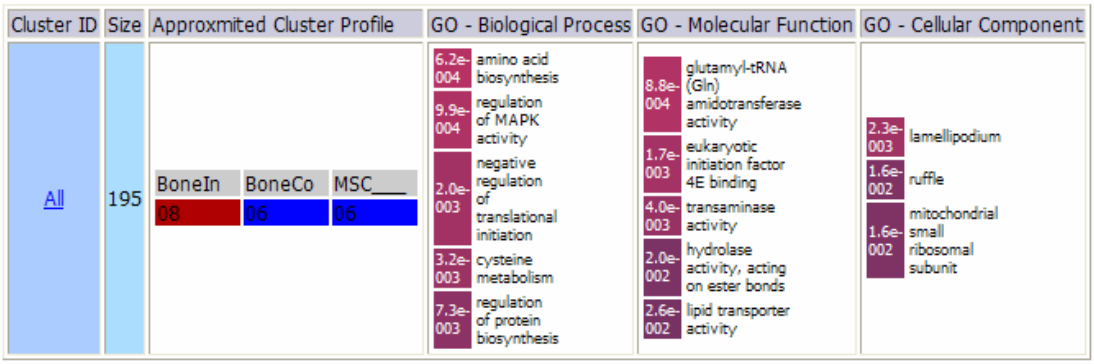


Table 8. Gene ontology enrichment test results. Top significant gene ontology terms for the 195 probesets over-expressed upon induction to bone.

ID	pValue	GO PS in cluster	Total GO PS on chip	Percentage	GO ID	GO Term
1	6.21e-004	4	22	18%	8652	amino acid biosynthesis
2	9.94e-004	2	3	67%	43405	regulation of MAPK activity
3	1.96e-003	2	4	50%	45947	negative regulation of translational initiation
4	3.23e-003	2	5	40%	6534	cysteine metabolism
5	7.25e-003	3	22	14%	6417	regulation of protein biosynthesis
6	8.58e-003	12	297	4%	6915	apoptosis
7	1.27e-002	6	105	6%	8284	positive regulation of cell proliferation

Gene ontology enrichment was calculated for the set of 195 probesets; Complete results (enriched GO terms and full gene list) are published online at http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_up/3_results.html.

After standardization, the 195 probe-sets were clustered using the SPC algorithm. Produced clusters were also tested for gene ontology enrichment. GO enrichment test results and complete gene list are published online at http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_up/3_c_summary.html.

Prominent significant GO terms and included genes:

- *Biological process – amino acid biosynthesis*

Gene Symbol	Gene Title
BCAT2	branched chain aminotransferase 2, mitochondrial
ASNS	asparagine synthetase
CBS	cystathionine-beta-synthase
PSAT1	phosphoserine aminotransferase 1

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_up/20060504T033433_GO_PS_8652.html

- *Biological process – apoptosis* *(6th on the biological process significant terms with a p-value of 8e-003)*

Gene Symbol	Gene Title
SQSTM1	sequestosome 1
LTBR	lymphotoxin beta receptor (TNFR superfamily, member 3)
GULP1	GULP, engulfment adaptor PTB domain containing 1
GADD45B	growth arrest and DNA-damage-inducible, beta
TNFRSF10B	tumor necrosis factor receptor superfamily, member 10b
GADD45B	growth arrest and DNA-damage-inducible, beta
CFLAR	CASP8 and FADD-like apoptosis regulator
FOXO3A	forkhead box O3A
CFLAR	CASP8 and FADD-like apoptosis regulator
BCL2L1	BCL2-like 1
TRIB3	tribbles homolog 3 (Drosophila)
HIPK2	homeodomain interacting protein kinase 2
CIDEC	cell death-inducing DFFA-like effector c
ELMO2	engulfment and cell motility 2 (ced-12 homolog, C. elegans)

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_up/20060504T033932_GO_PS_6915.html

Marked in yellow are ANTI-APOPTOSIS genes.

- Biological process – *cell differentiation*

Gene Symbol	Gene Title
SQSTM1	sequestosome 1
IFRD1	interferon-related developmental regulator 1
EFNB2	ephrin-B2
FRZB	frizzled-related protein
ANGPT1	angiopoietin 1
GADD45B	growth arrest and DNA-damage-inducible, beta
GPM6B	glycoprotein M6B
CSF1	colony stimulating factor 1 (macrophage)

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_up/20060510T115554_GO_PS_30154.html

Turning now to identifying genes that are **down-regulated** upon induction to bone, we have filtered in profiles whose expression Bone-Control is equal or lower compared to MSC, and lower on Bone-Induction compared to Bone-Control. These criteria yielded 36 profiles, representing 107 probe-sets, which we have used for gene ontology enrichment test.

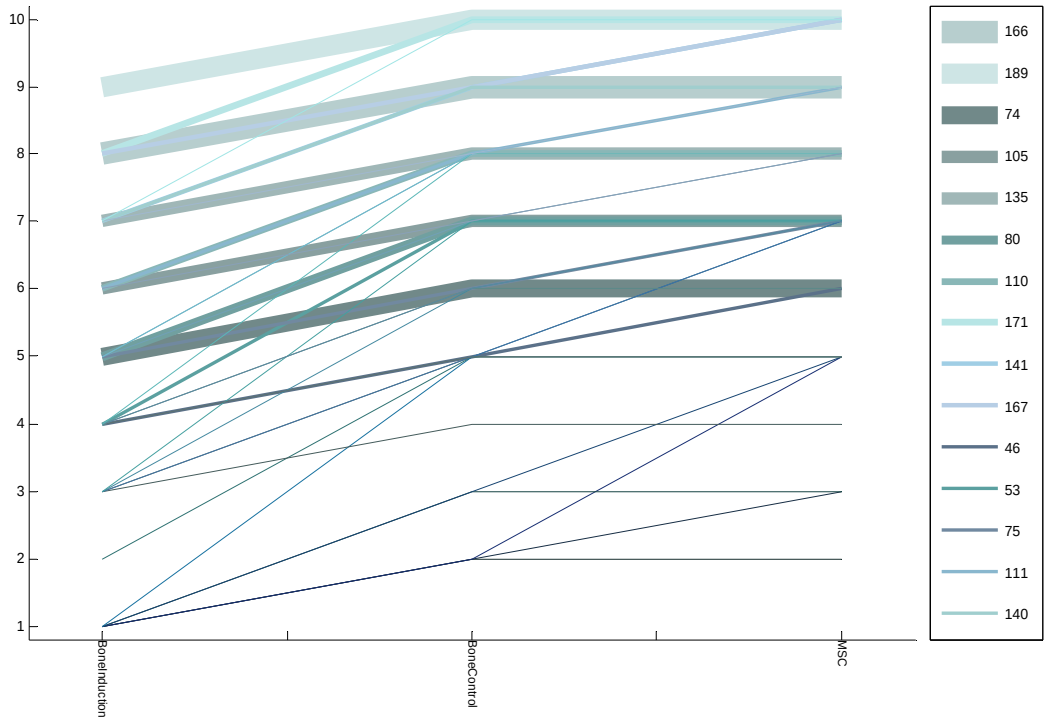


Figure 22. The 36 filtered profiles that are decreading from MSC to Bone-Induction. These profiles encapsulates 107 genes that are down-regulated upon induction to fat.

A. Un-standardized expression matrix

B. Standardized expression matrix

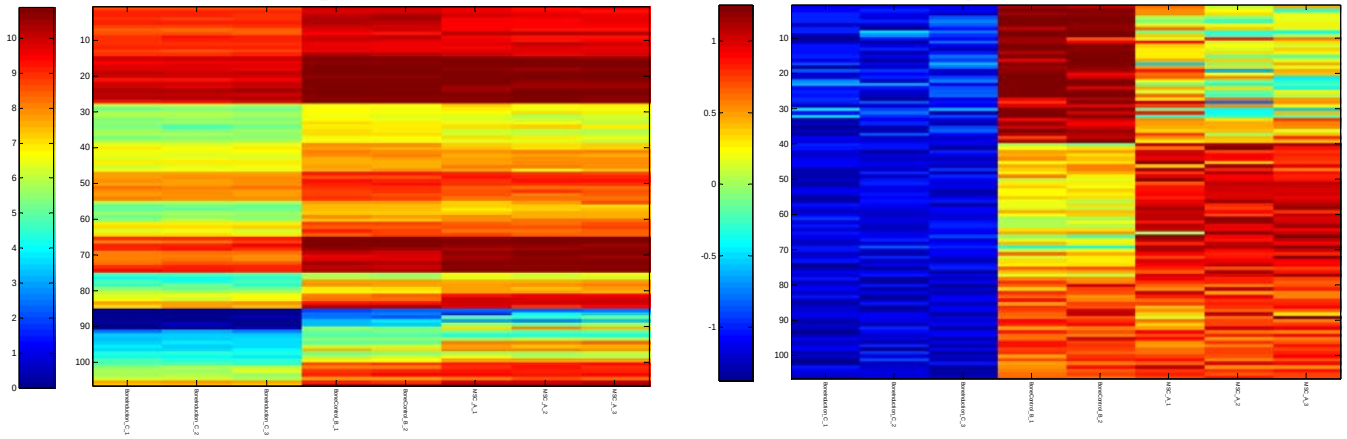


Figure 23. Expression matrices of the 107 probesets whose expression profile is decreasing from MSC to Bone-Induction. (A) Un-standardized expression matrix, probe-sets are ordered according to the 36 profiles. (B). Standardized expression matrix, probe-sets are ordered by SPC clustering.

Cluster ID	Size	Approxmited Cluster Profile	GO - Biological Process	GO - Molecular Function	GO - Cellular Component
All	106	BoneIn 06 / BoneCo 07 / MSC___ 07	9.4e-004 protein metabolism 1.9e-002 cell cycle arrest 2.1e-002 response to wounding 2.4e-002 phosphate transport 3.0e-002 calcium ion homeostasis	2.3e-004 procollagen-lysine 5-dioxygenase activity 2.8e-004 L-ascorbic acid binding 2.8e-004 oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors 5.1e-004 extracellular matrix structural constituent 2.1e-003 protein kinase C binding	1.2e-003 collagen 1.5e-003 collagen type IV 3.3e-003 basement membrane 1.4e-002 extracellular matrix (sensu Metazoa) 4.1e-002 Golgi membrane

Table 9. Gene ontology enrichment test results. Top significant gene ontology terms for the 107 probesets under-expressed upon induction to bone.

Gene ontology enrichment was calculated for the set of 107 probesets; Complete results (enriched GO terms and full gene list) are published online at http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_down/4_results.html.

After standardization, the 107 probe-sets were clustered using the SPC algorithm. Produced clusters were also tested for gene ontology enrichment. GO enrichment test results and complete gene list are published online at http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_down/4_c_summary.html.

Prominent significant GO terms and included genes:

- *Biological process – protein metabolism*

Gene Symbol	Gene Title
PLOD1	procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 1
LEPREL2	leprecan-like 2

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_down/20060504T035837_GO_PS_19538.html

- *Biological process – cell cycle arrest*

Gene Symbol	Gene Title
KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1 /// KH domain containing, RNA binding, signal transduction associated 1

CDKN2C	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)
MACF1	microtubule-actin crosslinking factor 1

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_down/20060504T035559_GO_PS_7050.html

- Cellular component – collagen

Gene Symbol	Gene Title
COL4A2	collagen, type IV, alpha 2
COL4A1	collagen, type IV, alpha 1
COL5A1	collagen, type V, alpha 1

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_down/20060504T040355_GO_PS_5581.html

- Cellular component - extracellular matrix (sensu Metazoa)

Gene Symbol	Gene Title
LAMA4	laminin, alpha 4
CSPG2	chondroitin sulfate proteoglycan 2 (versican)
CHI3L1	chitinase 3-like 1 (cartilage glycoprotein-39)
COL4A2	collagen, type IV, alpha 2
COL4A1	collagen, type IV, alpha 1
COL5A1	collagen, type V, alpha 1
CSPG2	chondroitin sulfate proteoglycan 2 (versican)

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI1/bone_down/20060504T040603_GO_PS_5578.html

Clustering analysis

Clustering Genes over Embryonic and Mesenchymal Samples

We have used the CTWC (Coupled Two-Way Clustering) [13] algorithm to cluster the 7,000 most variable probe-sets over all six sample types. Clustering was used to identify main data signals by putting together genes with similar expression patterns. Dataset preprocessing included setting a threshold of 20, log2 transformation and row standardization. The clustering operation was applied on the standardized dataset using default CTWC parameters, and yielded 43 stable clusters. Looking at CTWC output (Figure 24), it is apparent that when clustering the genes, the clustering algorithm first partitioned the genes into two large groups based on their expression on the three samples of the embryonic stem cells, and then further partitioned the data within these two large clusters, according to the expression patterns of the mesenchymal samples.

We have then tested each cluster for gene ontology enrichment in three classes of GO terms: Biological process, Molecular function and Cellular component. The calculated p-values express enrichment significance based on the hyper-geometric function. Enrichment tests were conducted on the gene level (rather than on the probe-set level). A GO term was called significant for a given cluster if it appears on at least two cluster genes that belong to the cluster and if its associated p-value was smaller than 0.01.

Full results including complete gene lists and GO enrichment analysis are available online at

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI2_ABCDEF7000/20060506T210124_CTWC_summary.html

A similar analysis was conducted on the five mesenchymal sample types without the ESC samples. Complete analysis results are available online at

http://bioinfo2.weizmann.ac.il/~netanely/Thesis/SI3_ABCDE7000/20060507T225315_CTWC_summary.html

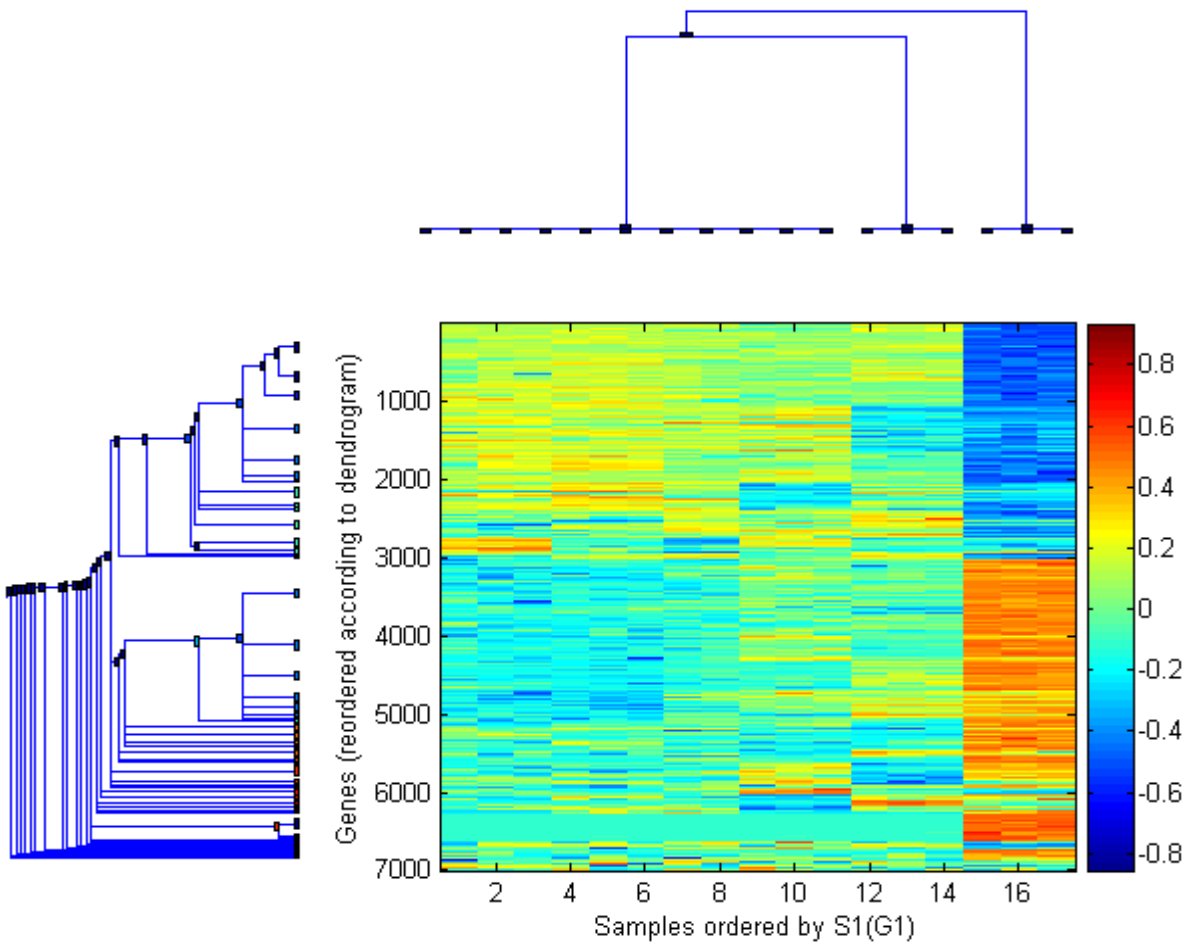
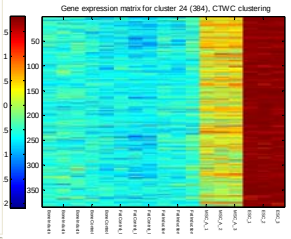
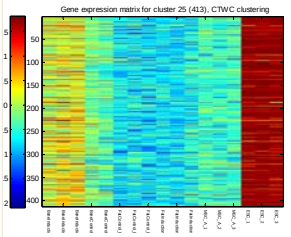
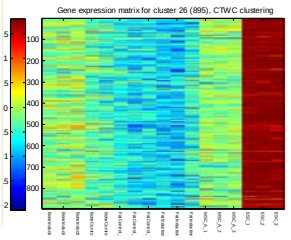


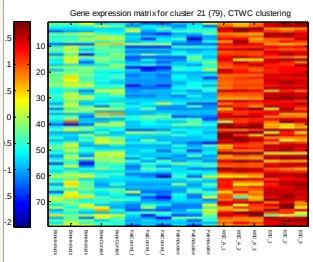
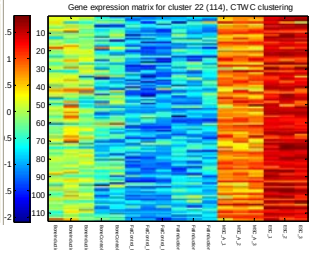
Figure 24. CTWC clustering output. The expression matrix was reordered; rows were reordered according to gene clustering and columns were reordered according to sample clustering. The dendrogram on the left represents gene clustering. The dendrogram on top represent sample clustering.

Selected clusters of interest are displayed below; each table row represents one cluster. The first column represent the cluster ID, the second column represent the cluster size. Approximated cluster profile based on dividing the normalized averaged cluster expression value to 10 levels is displayed on the third column. The profile is used as a concise way to display cluster expression pattern (It is only used for display; the clusters were determined by CTWC). The last three columns represent enrichment p-values for the most significant GO terms of the three GO classes: Biological process, Molecular function and Cellular component.

- Clusters over-expressed in ESCs are enriched with 'Biological process' GO terms related to *mitosis*, *cell-cycle*, *DNA replication* and *mRNA processing*. In the 'Molecular function' class, they are enriched mainly in *nucleic acid binding* terms, and in the 'Cellular component' class they are enriched in the *nucleus*.

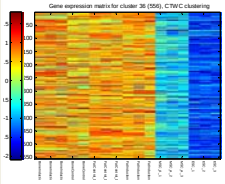
Cluster ID	Size	Approxmited Cluster Profile	GO - Biological Process	GO - Molecular Function	GO - Cellular Component	
24	384	BoneIn 04 / BoneCo 04 / FatCon 04 / FatInd 04 / MSC___ 06 / ESC___ 10	4.5e-013 cell cycle 4.0e-012 DNA replication 1.2e-011 mitosis 3.5e-011 cell division 1.9e-009 DNA replication initiation	1.8e-008 DNA-dependent ATPase activity 3.1e-008 nucleotide binding 1.3e-006 RNA binding 6.4e-006 ATP binding 5.9e-004 DNA bending activity	1.2e-006 nucleus 2.6e-004 kinetochore 4.6e-004 chromatin 8.7e-004 condensed chromosome 9.3e-004 spindle	
25	413	BoneIn 06 / BoneCo 05 / FatCon 04 / FatInd 04 / MSC___ 04 / ESC___ 10	6.9e-005 mRNA processing 1.1e-004 chromosome organization and biogenesis 5.1e-004 RNA splicing 7.9e-004 mRNA export from nucleus 1.5e-003 response to DNA damage stimulus	1.1e-005 nucleic acid binding 3.0e-005 helicase activity 2.6e-004 RNA binding 2.9e-004 nucleotide binding 1.3e-003 peptide binding	7.5e-009 nucleus 4.0e-003 nucleoplasm 8.8e-003 centriole 2.1e-002 condensed chromosome 2.3e-002 DNA-directed RNA polymerase II, core complex	
26	895	BoneIn 05 / BoneCo 04 / FatCon 04 / FatInd 04 / MSC___ 05 / ESC___ 10	0.0e+000 mRNA processing 6.9e-009 nuclear mRNA splicing, via spliceosome 8.8e-009 DNA repair 3.5e-008 response to DNA damage stimulus 8.5e-007 RNA processing	1.3e-011 RNA binding 3.9e-011 nucleic acid binding 1.2e-005 pseudouridylate synthase activity 7.6e-005 binding 1.0e-004 nucleotide binding	0.0e+000 nucleus 2.1e-007 nuclear pore 4.5e-006 heterogeneous nuclear ribonucleoprotein complex 1.4e-005 nucleoplasm 1.9e-004 chromatin	

- Similarly, clusters exhibiting over-expression in both ESC and MSC samples, were found to be enriched with mitosis, cell-cycle and spindle-organization.

Cluster ID	Size	Approxmited Cluster Profile	GO - Biological Process	GO - Molecular Function	GO - Cellular Component	
21	79	BoneIn 04 / BoneCo 04 / FatCon 03 / FatInd 03 / MSC___ 08 / ESC___ 09	3.7e-009 mitosis 1.4e-005 spindle organization and biogenesis 3.0e-005 cell cycle 4.7e-005 mitotic sister chromatid segregation 3.8e-004 cell division	4.9e-005 ribonucleoside-diphosphate reductase activity 4.0e-003 structural constituent of cytoskeleton 5.8e-003 microtubule motor activity 6.1e-003 chromatin binding 8.6e-003 cysteine protease inhibitor activity	1.1e-005 chromosome, pericentric region 6.0e-004 microtubule 2.8e-003 nucleosome 4.7e-003 spindle 8.8e-003 chromosome	
22	114	BoneIn 05 / BoneCo 04 / FatCon 03 / FatInd 03 / MSC___ 07 / ESC___ 09	1.9e-006 mitosis 4.6e-005 spindle organization and biogenesis 1.6e-004 phosphoinositide-mediated signaling 5.5e-004 cell division 1.7e-003 DNA recombination	3.7e-003 exonuclease activity 4.3e-003 RNA binding 7.2e-003 nuclease activity 2.3e-002 nucleotide binding 2.6e-002 ligase activity	7.1e-003 chromosome 1.0e-002 spindle 1.5e-002 nucleus 1.7e-002 kinetochore 2.1e-002 mediator complex	

- Exhibiting an opposite profile, a gene cluster under-expressed in ESC was found to be enriched with cell-matrix adhesion. The cluster includes probe-sets whose expression is elevated along the differentiation pathway.

Cluster ID	Size	Approxmited Cluster Profile	GO - Biological Process	GO - Molecular Function	GO - Cellular Component
36	556	BoneIn 07 | BoneCo 07 | FatCon 07 | FatInd 07 | MSC___ 04 | ESC___ 02	1.5e-004 cell-matrix adhesion 1.3e-003 cAMP biosynthesis 1.9e-003 positive regulation of angiogenesis 4.5e-003 angiogenesis 6.6e-003 regulation of MAPK activity	4.8e-004 transcription corepressor activity 5.9e-003 MP kinase activity 6.0e-003 phosphorus-oxygen lyase activity 6.2e-003 adenylate cyclase activity 6.2e-003 extracellular matrix structural constituent	7.6e-003 lipid particle 8.9e-003 ATP-binding cassette (ABC) transporter complex 1.6e-002 lysosomal membrane 1.7e-002 endosome 2.4e-002 neuron projection

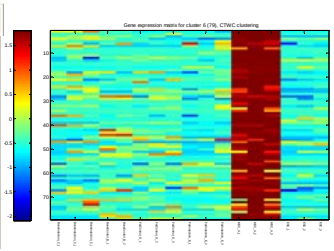


The following are among the cell-matrix adhesion genes included in the cluster:

Gene Symbol	Gene Title	Pathway
ITGB5	integrin, beta 5	Integrin-mediated_cell_adhesion
NID1	nidogen 1	---
PKD1	polycystic kidney disease 1 (autosomal dominant)	---
SGCE	sarcoglycan, epsilon	---
TNXB	tenascin XB	---
ECM2	extracellular matrix protein 2, female organ and adipocyte specific	---
C9orf127	chromosome 9 open reading frame 127	---
ITGA7	integrin, alpha 7	Integrin-mediated_cell_adhesion
SNED1	sushi, nidogen and EGF-like domains 1	---
ITGA8	integrin, alpha 8	Integrin-mediated_cell_adhesion
ADAM15	ADAM metallopeptidase domain 15 (metargidin)	---

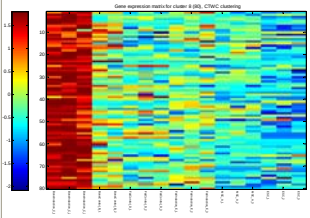
- Probe-sets exclusively over-expressed in MSC were found to be enriched with *immune* related GO terms such as *immune-response*, *inflammatory response* and *chemotaxis*. *Chemokine* and *cytokine* activity are among the enriched GO terms in the 'Molecular function' class.

Cluster ID	Size	Approxmited Cluster Profile	GO - Biological Process	GO - Molecular Function	GO - Cellular Component
6	79	BoneIn 05 | BoneCo 05 | FatCon 05 | FatInd 04 | MSC___ 10 | ESC___ 04	0.0e+000 immune response 4.0e-007 response to virus 4.6e-006 inflammatory response 9.9e-006 chemotaxis 3.9e-005 cell-cell signaling	2.4e-008 chemokine activity 1.5e-006 cytokine activity 5.5e-005 growth factor activity 2.9e-004 collagenase activity 9.9e-003 heparin binding	1.2e-006 extracellular region 2.4e-006 extracellular space 1.3e-003 lysosome



- Probe-sets exclusively over-expressed in differentiated bone samples (Bone-Induction) were found to be enriched with *activation of NF-kappaB-inducing kinase, mesoderm development* and *apoptosis*.

Cluster ID	Size	Approxmited Cluster Profile	GO - Biological Process	GO - Molecular Function	GO - Cellular Component
8	80	BoneIn 10 / BoneCo 06 / FatCon 05 / FatInd 05 / MSC___ 04 / ESC___ 04	2.4e-003 activation of NF-kappaB-inducing kinase 1.1e-002 mesoderm development 1.4e-002 protein biosynthesis 2.1e-002 tRNA aminoacylation for protein translation 2.4e-002 apoptosis	2.3e-003 protein self binding 1.5e-002 double-stranded RNA binding 2.3e-002 tRNA ligase activity 3.5e-002 electron transporter activity 4.8e-002 transmembrane receptor activity	4.4e-002 cytoskeleton 4.9e-002 Golgi apparatus

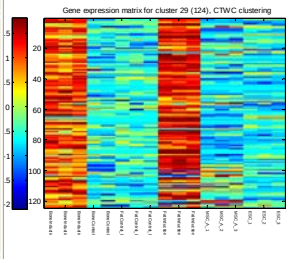


Apoptosis related genes found in this cluster:

Gene Symbol	Gene Title	Pathway
PPP1R15A	protein phosphatase 1, regulatory (inhibitor) subunit 15A	---
MAP3K5	mitogen-activated protein kinase kinase kinase 5	---
BCL2L1	BCL2-like 1	Apoptosis // GenMAPP /// Apoptosis_GenMAPP // GenMAPP /// Apoptosis_KEGG // GenMAPP
TNFRSF10B	tumor necrosis factor receptor superfamily, member 10b	Apoptosis // GenMAPP
YARS	tyrosyl-tRNA synthetase	Phenylalanine, tyrosine and tryptophan biosynthesis // KEGG /// Aminoacyl-tRNA synthetases // KEGG
HIPK2	Homeodomain interacting protein kinase 2	---

- The following gene cluster includes probe-sets that are highly expressed only in the two differentiated samples type (Bone-Induction and Fat-Induction). The genes were found to be enriched with *actin filament-based movement*.

Cluster ID	Size	Approxmited Cluster Profile	GO - Biological Process	GO - Molecular Function	GO - Cellular Component
29	124	BoneIn 08 / BoneCo 04 / FatCon 04 / FatInd 08 / MSC___ 04 / ESC___ 04	2.3e-003 actin filament polymerization 3.9e-003 N-acetylglucosamine metabolism 3.9e-003 sulfur metabolism 7.0e-003 axonogenesis 7.2e-003 fatty acid biosynthesis	2.0e-009 cadmium ion binding 1.1e-005 copper ion binding 6.0e-004 N-acetylglucosamine 6-O-sulfotransferase activity 6.1e-003 lipid transporter activity 7.3e-003 mRNA binding	1.9e-003 intrinsic to Golgi membrane 2.6e-002 extracellular matrix (sensu Metazoa) 4.6e-002 basement membrane



Summary and Discussion

Differentiation is a very important biological process by which general progenitor cells give rise to specialized mature cells by acquiring specific functions. Though most cells in an adult multi-cellular organism contain the entire genome, different mature cells express different sub-sets of the cell genes in varying levels of expression. As cells differentiate, their transcriptional program changes, different pathways of differentiation include convergence to specific mature transcriptional programs. Little is know about the genetic mechanisms that govern the ability of both embryonic and adult stem cells to divide indefinitely in culture and yet retain their ability to give rise to many different types of mature cell through differentiation [45].

In this work, we have studied gene expression profiles underlying the differentiation process of embryonic and mesenchymal stem cells into differentiated bone and fat cells. Mesenchymal stem cells were derived from donors undergoing orthopedic surgery and were induced in-vitro to differentiate into mature bone and fat cells. The expression of 14,500 genes was measured using DNA microarrays on 6 sample types, representing different stages along the differentiation pathway. Our dataset included the following sample types: Embryonic stem cells, mesenchymal stem cells, Bone-Control, Bone-Induction, Fat-Control and Fat-Induction. The data was analyzed using various computational methods including statistical tests (t-test, rank-sum, FDR, ANOVA), clustering algorithms (hierarchical clustering and SPC) and expression profiling. An additional pivotal component of our analysis used the gene ontology (GO) annotation system to test for over-representation of GO terms in gene lists produced by the mentioned statistical and clustering methods.

The most prominent signal identified in the data by various methods, relates to the distinction of the embryonic stem cell samples compared with all other mesenchymal stem cell and differentiated samples. ESC samples were found to both over-express and under-express hundreds of probe-sets compared with the

mesenchymal samples. This observation was detected by several analysis methods including global dataset expression histograms, clustering, profiling and comparative analysis of differentially expressed genes. In the latter, out of 12,461 differentially expressed genes identified by ANOVA over the 6 sample types, there were 1,678 more over-expressed than under-expressed probe-sets when comparing embryonic stem cells to mesenchymal stem cells. ESCs also over-expressed 916 - 1286 more probe-sets compared with the other mesenchymal samples. This observation is consistent with previous observations made by others [23, 40] and with the "Just In Case" theory (by which stem cells promiscuously express genes which later shut down upon differentiation). This transcriptional program may be attributed to evidence indicating that much of the chromatin of embryonic and adult stem cells is in an open, accessible state, which might allow the promiscuous expression of lineage-specific genes [23].

Interestingly, MSC samples were not found to exhibit marked over-expression of genes compared with both differentiated bone and fat samples (MSC express 717 more over-expressed probe-sets compared to Fat-Induction and 1376 less over-expressed probe-sets compared to Bone-Induction). In fact, several analysis methods revealed that the differentiated bone samples (Bone-Induction) highly express many more genes than the mesenchymal stem cells. This observation may be ascribed to the genetic variance of the three donors from which different mesenchymal samples were derived, and requires further investigation.

In order to identify genes whose expression changes significantly upon induction to bone or fat, we have used t-test to compare the Control and Induction samples of bone and fat. We then filtered out genes whose expression profile from MSC to Induction is non-monotonic, in order to compensate for the donor variability problem. Gene annotation enrichment tests based on Gene ontology terms were conducted (separately) on the resulting genes.

As expected, the 210 genes up-regulated in differentiated fat included *fat metabolism* genes (ACACB, SCD, PTGS1) and *lipid transporter activity* genes

(APOD, APOE, APOL1). Several of the identified up-regulated genes have been previously linked to fat-differentiation by others [46]: PPAR-γ (which when activated, promotes adipogenesis and inhibits osteogenesis) [46, 47], Collagen type XI (COL11A1), alcohol dehydrogenase IB (ADH1B), IGF2, IMPA2 and APOE. A significant number of genes from that group are associated with the *extra-cellular matrix*.

However, the 195 genes up-regulated upon differentiation to bone were not found to be enriched with GO terms that are directly related to osteogenesis. 'Biological Process' GO terms such as *Apoptosis (*12 genes, including anti-apoptotic genes like BCL2L1, CFLAR and FOXO1A) and *cell-differentiation* (8 genes, including CSF1, GADD45B, FRZB) were identified as significantly enriched. Enrichment with *apoptosis* genes may be explained by the natural role of apoptosis in the building of bone; in-vivo, the bone undergoes continuous remodeling in which osteoclasts resorb aged or damaged bone, leaving space for osteoblasts to make new bone [48]. An alternative explanation may rely in the culturing conditions that were used to grown the differentiated bone samples. Culturing of bone induction samples started at low density and the cells were cultured for several days until they filled the plate (unlike induction of fat samples, which started at full plate). We hypothesize that the fact that the induced bone cells reached spatial limit few days before RNA extraction may explain the over representation of apoptosis related GO terms in the list of differentially expressed genes. Further analysis is required to explain this observation.

The 107 genes down-regulated in differentiated bone compared to MSC and Bone-Control were found to be enriched with collagen (types IV and V).

Clustering analysis conducted on the 7000 most variable dataset probe-sets using the SPC algorithm yielded 43 stable clusters. Again, the most prominent signal was linked to the embryonic stem cell samples. Clusters containing genes that are over expressed in ESCs, exhibited a high statistically significant

enrichment with GO terms such as *cell-cycle*, *mitosis*, *DNA-replication*, cell-division and *DNA-repair* (SPC clusters 24, 25 and 26). A significant enrichment with GO terms like *mitosis*, *cell-cycle* was also found in clusters whose genes are over-expressed in both ESCs and MSCs (clusters 21 and 22). This observation demonstrates that unlike differentiated mature cells, stem cells are still actively cycling through the cell cycle and have a proliferative capacity. The enrichment of ESCs with GO terms such as '*DNA-repair'* and '*response to DNA damage stimulus'* are consistent with recent studies emphasizing the importance of DNA repair in stem cells. In order to maintain the pool size of the long-lived stem cells, stem cells cannot promiscuously use apoptosis (like the more disposable mature cells) to control DNA damage, and thus DNA repair mechanisms are of an extreme importance in stem cells [49].

The genes of cluster 24 (containing 384 probe-sets) are a good example for genes whose expression decreases along the differentiation pathway. These genes exhibit a very high relative expression on ESCs, high but decreased expression on MSCs and low expression on all bone and fat samples. This cluster includes 43 'cell-cycle' genes, 24 'DNA-repair' and 20 'mRNA processing' genes (groups may overlap).

129 genes over-expressed only in both differentiated samples (Bone-Inducion and Fat-Induction) are included in cluster 29: BMP6 (bone morphogenetic protein 6), OMD (osteomodulin), ADH1B (alcohol dehydrogenase IB), PPARG (peroxisome proliferative activated receptor, gamma), FRZB (frizzled-related protein) and others. These genes may be involved in pathway shared by the two differentiation pathways, as many of them monotonically increases along the differentiation pathway.

Regarding enrichment of terms from the 'Cellular Component' GO category, it is worth mentioning that almost all clusters exhibiting over-expression in ESC samples are associated with the *nucleus*, whereas clusters over-expressed in bone or fat samples are associated with the *extra-cellular matrix*.

The analysis conducted in this work identified and classified many genes whose expression changes along the pathway of mesenchymal differentiation. Gene ontology was used to organize and interpret the relevance of identified genes to the biological processes underlying differentiation to bone and to fat. The analysis has focused on the level of gene groups due to the high complexity and size of the data, and additional investigation is needed at the gene level.

Donor variance and differences in sample culturing challenged our ability to detect differentially expressed genes whose expression changes due to the actual investigated process of differentiation. Future microarray experiments used to investigate similar processes may benefit greatly from reducing the background noise, perhaps by deriving investigates samples from mice of the same strain.

Part 3

Leukemic Over Expression of Tissue Specific Genes

Collaboration with Prof. Leo Sachs and Dr. Joseph Lotem

Biological Background

General Introduction to Cancer

Cancer develops through a multi-step process by which normal cells transform into malignant cells due to genetic and epigenetic changes. Each such decisive step provides the transformed cell with a certain survival and growth advantage enabling it to over-grow its surrounding normal cells, eventually spreading uncontrollably through the body – severely hampering its normal physiology.

Hanahan at al. enumerated four traits that must be acquired by transformed cells during tumorigenesis: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis) and limitless replicative potential. These traits are essential for the central characteristic of cancer: uncontrolled proliferation. Two additional hallmarks are needed to turn the cancer into a killer of the organism: sustained angiogenesis, and tissue invasion and metastasis [50].

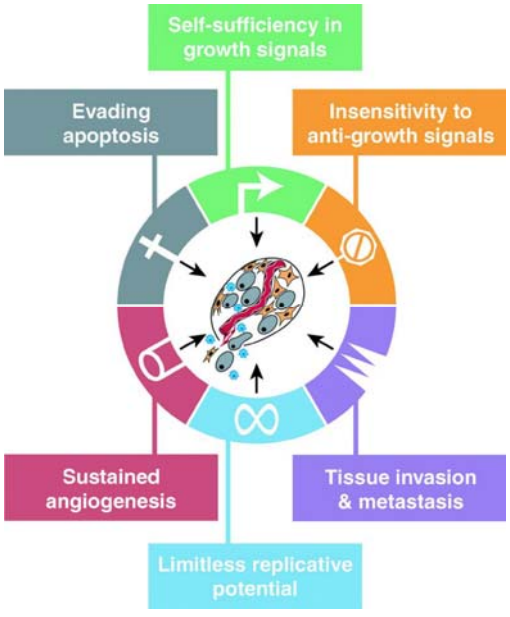


Figure 1. Cancer acquired capabilities.

Mutations, deletions and chromosomal translocations are genetic mechanisms altering normal gene function, occasionally promoting tumorigenesis by giving rise to oncogenes (dominant gain-of-function genes promoting the development of cancer) and by disrupting the anti-malignant function of tumor suppressor genes (recessive loss-of-function genes). Epigenetic changes like DNA methylation and histone modification also play a major role in cancer development through transcription silencing of tumor suppressor genes and/or activation of oncogenes [51] .

The genetic and epigenetic alterations that lead to cancer disrupt the transcriptional program of transformed cells, thus affecting *proliferation*, *differentiation* and *apoptosis*, which are pivotal in determining cell population development and net growth. During differentiation, general primitive progenitor cells become committed to a specific function. The progenitors' proliferative potential is significantly reduced during differentiation because differentiation implies a definite withdrawal from the cell-cycle. When such a cell becomes malignant, deranged cellular regulation may cause differentiation block. When the cell's apoptosis mechanism is also impaired and therefore cannot avoid the propagation of damaged DNA to progeny cells, the transformed cell may keep proliferating in its undifferentiated state [52]. Differentiation arrest is therefore an important component in the pathogenesis of many cancers; here we focus on Acute Myeloid Leukemia to demonstrate the involvement of differentiation arrest in tumorigenesis.

Arrest of Differentiation and Tumorigenesis – AML as an Example

The term leukemia refers to cancers of the white blood cells. Leukemia is a very heterogeneous disease, composed of many subtypes. In general, leukemias are classified into **acute** (rapidly developing) and **chronic** (slowly developing) forms. Leukemia is also divided by type of white blood cell that is affected: **ALL** (Acute Lymphoid Leukemia) and **AML** (Acute Myeloid Leukemia). ALL mainly affects children and young adults whereas AML mainly affects adults with

increasing frequency with age. AML is more difficult to treat in comparison to ALL; overall five years survival rates for AML remain below 60% [53] .

In AML, the malignant myeloid cells, called myeloblasts, fail to mature into different types of white blood cells. The myeloblasts proliferate rapidly, accumulate in the bone marrow, depriving the healthy blood cells of resources, and eventually spread into the bloodstream and other vital organs. The lack of the various types of healthy blood cells results in symptoms such as anemia, abnormal bleeding and infections; lack of functioning blood cells leads to death [54].

AML is an excellent model for studying the relationship between differentiation regulation and cancer progression for several reasons. First – technical: malignant leukemic cells in the blood stream are easily accessible and there are many available isolation protocols to obtain relatively pure hematopoietic sub-populations of cells. Second, the blood system in general and hematopoietic differentiation in particular, is among the most extensively studied and best understood systems. In recent years, myeloid lineage-specific transcription factors were identified, and their specific role in differentiation were established. These transcription factors regulate differentiation by several mechanisms such as activation of lineage specific genes, inhibition of alternative pathways, inhibition of proliferation, and induction of apoptosis. Specific cell fate commitment in the hematopoietic system is determined by alternative expression of specific combinations of such transcription factors [55].

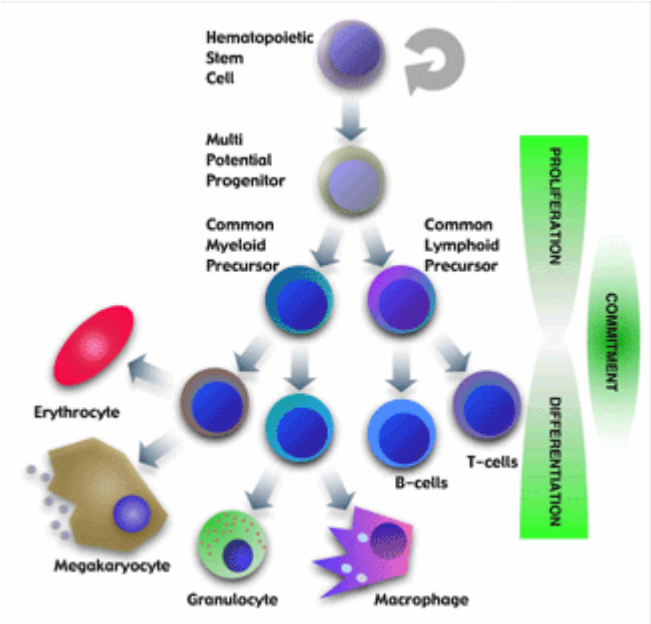


Figure 2: Proliferation versus differentiation in hematopoietic differentiation. The commitment process is characterized by massive cell proliferation in the early phase followed by successive restriction to distinct cell lineages and to cell differentiation.

Functional inactivation of the differentiation inducing transcription factors mentioned above may lead to differentiation block, in which hematopoietic precursors remain in their undifferentiated state, while retaining proliferative capacity. Normally, as differentiation progresses and cells mature, their proliferative potential is reduced and eventually even lost altogether. The stage at which differentiation is arrested depends on the pathway that was disrupted. Moreover, many AML subtypes are now associated with specific genetic lesions – mutations or chromosomal translocations. These genetic lesions may enable genetic classification of AML sub-types that would replace the present phenotype based French-American-British (FAB) classification of AML [56].

For example, AML subtype M3 (APL - Acute promyelocytic leukemia) is associated in most cases with a translocation between chromosomes 15 and 17. This translocation may give rise to a fusion protein called PML-RARa, which is the result of fusing the promyelocytic leukemia (PML) gene on chromosome 15, and the retinoic acid receptor alpha (RARα) gene on chromosome 17. The PML gene

encodes a growth suppressing transcription factor, and the RARα gene regulates myeloid differentiation. The chimeric fusion protein arrests the myeloid cell maturation at a specific differentiation step - the promyelocytic stage, and this leads to the increased proliferation of promyelocytes [57].

Differentiation Therapy

Treatment of several cancer types, such as APL (AML Subtype M3) can now benefit from an emerging therapeutic field called 'differentiation therapy'. The underlying strategy is to reverse (using chemical means) the cell's transition to the malignant state by releasing the differentiation block. Most APL patients are now successfully treated with all-trans-retinoic acid (ATRA) which activates the retinoid receptor RAR, causes degradation of the oncogenic PML-RARα fusion protein and re-enables differentiation of the malignant promyelocytes', thus decreasing their proliferative potential. With the introduction of ATRA, APL has now become the most curable AML subtype [58].

Such clinical breakthroughs are based on studies like those conducted by L. Sachs et al. [59] where he has investigated whether malignant cells can revert back to cells that again show normal growth control. L. Sachs and J. Lotem [60] have used myeloid leukemic cells as a model system in order to determine whether malignancy can be suppressed by inducing differentiation with normal cytokines. They found that several myeloid leukemic cells could be induced both in vitro and in vivo, by adding different cytokines such as IL-6, IL-1, GM-CSF, G-CSF and IL-3, and various other compounds including retinoic acid, to differentiate to non-dividing mature granulocytes and/or macrophages.

A second example for using differentiation therapy methods is found in hepatocellular carcinoma (liver cancer). It is a common solid tumor, considered very hard to treat effectively. MYC (myelocytomatosis viral oncogene homolog) is an oncogene that was shown to be over expressed in many types of cancer. Shachaf et al. have demonstrated that inactivation of the MYC gene alone

suffices to induce regression of invasive liver cancers in mice. Within 4 days of MYC inactivation, the liver tumors differentiated into normal liver cells accompanied by apoptosis [61] [62].

Spira et al. [63] describe the general process by which 'differentiation therapy' is believed to work: "Although there are probably mechanistic differences in how the various agents lead to differentiation, the overall process itself is likely to function by allowing malignant tumor cells to revert to a more benign form, in which their replication rates are lower compared with malignant forms, leading to a decreased tumor burden. They might also have a decreased tendency for distant metastatic spread, and the process may also restore traditional apoptotic pathways, all of which could improve a patient's prognosis."

These and other examples demonstrate the promising potential of using 'differentiation therapy' to treat different types of cancer. By identifying and bypassing or correcting a specific malfunctioning element within the differentiation process, it may allow us to restore the non-malignant phenotype.

The prospects of using 'differentiation therapy' agents to cure cancer, especially when compared to current treatment methods, are great; examining the clinically tested example of using ARTA to treat APL leukemia reveals that it is by far more efficient than conventional treatments, cheaper, has less side effects and does not damage healthy tissues due to its high biological specificity [64].

Cancer and Stem Cells

Recent studies suggest that adult stem cells may play a key role in cancer. Adult stem cells are multipotent cells responsible for tissue renewal and repair of aged or damaged tissue. They are capable of self-renewal, which enables them to maintain their cell population size; they are also capable of differentiating into specific mature cells when needed, to regenerate the tissue they reside in. Adult stem cells divide asymmetrically - one daughter cell is a new stem cell and the second is a progenitor cell, which differentiates and proliferates into mature cells.

Adult stem cells were identified in many tissues, including skin, gut, brain, liver and mesenchyme (the latter stands at the center of the first part of this essay).

Cancer cells and stem cells share several common properties. In general, both are undifferentiated; they exhibit a general phenotype that is not committed to a specific function. Both may also exhibit elevated proliferative capacity. It is unsurprising therefore that the initial insight of researchers examining cancer tissue samples under a microscope, already in the mid 19th century, was to relate cancer cells to embryonic cells due to their histological resemblance [65]. Furthermore, it has been recently proven that tumors are heterogeneous, composed of both rare 'tumor initiating cells' and abundant 'non-tumor initiating cells'. The 'tumor initiating cells' were found to have self-renewal and proliferation ability, express typical markers of stem cells and are also resistant to drugs (which may explain why it is so difficult to eradicate tumors). 'Tumor initiating cells' were identified in leukemia and in various solid tumors including breast cancer. They were therefore named – 'Cancer stem cells' [66]. In addition, a number of factors that govern the fate of normal adult stem cells also play a role in malignant cell transformation, such as Wnt, Oct-4, Bmi-1 and Evi1 [67, 68] .

Since fully differentiated somatic cells are relatively short-lived, it is less likely that they would have the chance to accumulate the number of genetic and epigenetic changes needed to set off tumorigenesis. It is more likely that a long-lived cell, already capable of self-renewal (such as an adult stem cell), will be the target of such changes [69]. Therefore, according to the 'Cancer stem cell' theory, cancer stem cells originate from normal stem cell and/or from progenitor cells by mutation and epigenetic changes. Further mutations of these cells lead to formation of heterogeneous tumor containing different tumor cells. When such a tumor is treated by chemotherapy, most of the cells are killed, but cancer stem cells survive because of their higher resistance. These stem cells can initiate

malignant tumor growth sometimes after long period of time and can also be the source of metastasis spread of tumor cells [66].

According to an alternative (or complementary) explanation [70], cancer stem cells originate from mature somatic cells, which by mutation underwent *dedifferentiation* or reprogramming and regained stem-like properties, mainly self-renewal and elevated proliferation rate.

To summarize, the 'cancer stem cell' theory argues that cancer is simply a process if uncontrolled proliferation of abnormal adult stem cells that are unable to enter the pathway of terminal differentiation. According to this hypothesis, differentiation therapy is aimed at inducing cancer cells back into the natural pathway of terminal differentiation and eventual senescence.

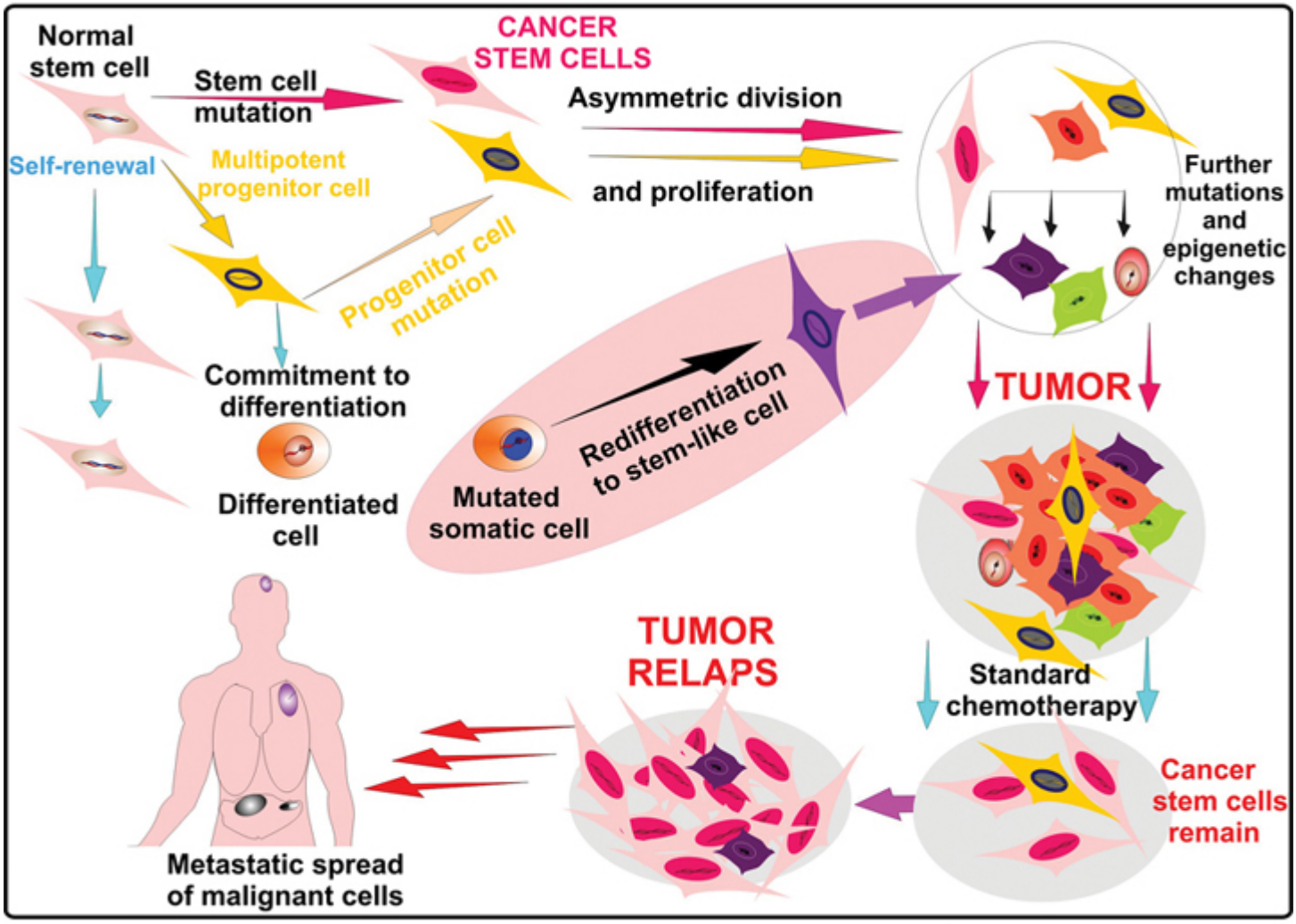


Figure 3. Origin of cancer stem cell and implementation to tumor therapy.

Adult Stem Cell Plasticity and the Prospects of 'Trans-Differentiation Therapy'

As described above, adult stem cells were identified in various body tissues and they are believed to be a source for specific mature cells needed to replenish the tissue within they reside, in response to injury or as part of normal physiology. Taking the hematopoietic stem cell as an example, it is a multi-potent adult stem cell, capable of giving rise to the many different mature cell types of the hematopoietic lineage. However, as has been recently demonstrated [26, 71],

under certain conditions, the progeny cells of adult stem cells are not restricted only to their own lineage, and can give rise to differentiated cells of a different lineage, a process called *trans-differentiation*. For example, bone marrow derived stem cells were shown to have the potential of differentiating to mature cells of the heart, liver, kidney, lungs, GI tract, skin, bone, muscle, cartilage, fat, endothelium and brain. A subpopulation of cells in the brain can differentiate into all of the major cell types in the brain and also into hematopoietic and skeletal muscle cells [72].

Considering the remarkable resemblance and shared properties of adult stem cells and cancer cells and based on the 'cancer stem cell' theory, we raise a hypothesis by which cancer may be treated by inducing it to trans-differentiate into mature cells other than its tissue of origin, and thus positive effects, similar to those gained by using 'differentiation therapy' may be achieved. Since global gene expression disturbance is associated with many cancers, certain trans-differentiation paths may be easier to pursue in a given malignant cell compared with the original path of differentiation that is blocked by possibly many impaired cell components. 'Trans-differentiation therapy' may therefore offer a way to by-pass impaired differentiation pathway by introducing alternative differentiation paths to mature cells of other lineages.

But which mature tissue is the most likely trans-differentiation target for a given cancer? In order to evaluate the possibility of trans-differentiating a given malignant cell to different types of mature tissues, we have decided to examine gene expression profiles of cancer cells and look for expression of genes that are specific to mature tissues different from the tissue from which the cancer originated. Identification of such genes may tell us about the linkage between the examined cancer and the normal mature tissue. Cancer cells that express genes specific to a different normal tissue may have extensive transcriptional accessibility to genes of that tissue, and therefore stand a higher chance of being successfully re-differentiated towards this mature tissue.

The Questions Posed

As discussed above, genetic and epigenetic changes contribute to the development of cancer and lead to abnormalities in regulation of cell viability, multiplication and differentiation. Clearly, such changes in DNA and chromatin structure are reflected in aberrant gene expression, and indeed, analysis of global gene expression detected both up-regulation and down-regulation of many genes in different types of cancer, when compared to their respective normal tissues of origin [73-78]. We pose here several questions regarding the genes whose expression has been significantly modified by the malignant transformation. First - *are the up-regulated genes in cancer cells limited only to those genes that are normally preferentially expressed in the same tissues from which the cancer originated?* We have previously shown [79] that cells of a mouse myeloid leukemia cell line highly express various genes that are normally preferentially expressed in different non-hematopoietic tissues including neuronal, liver, testis and muscle. We have now investigated whether this phenomenon is limited to a particular murine cell lines, or is it common to different types of human cancers? We addressed also another question: whether the tissue specific characteristics of the genes that are up-regulated in cancer are universal, or do they vary between different human cancer cell lines [80] and different subtypes of human leukemia from patients [76, 81].

Materials and Methods

Data sets

Three DNA microarray data sets were used: In the first data set, mRNA expression levels of genes in normal human tissues and in various human cancer cell lines [80] were measured using 2 DNA microarrays, the Affymetrix HG-U133A array and the GNF1H, a custom designed array [80]. The data set downloaded from Su et al. (http://wombat.gnf.org/index.html) contained 33,689 probe sets (PS). We removed all PS that were mapped to more than one gene symbol, leaving 33,440 PS that were used for further analysis. The downloaded data set included 72 normal human tissue samples in duplicates and 7 human cancer cell lines also in duplicates [80]. The cancer cell lines [80] included the T cell lymphoma MOLT4, the B-cell lymphoma 721, the Burkitt's lymphomas Raji and Daudi, the myeloid leukemia HL-60, the chronic myeloid leukemia derived cell line K562 and the colorectal carcinoma SW480.

The two other data sets used included mRNA expression data from leukemic blast cells of 132 pediatric patients with different acute lymphoid leukemia (ALL) subtypes [81], 5 pediatric patients with T-ALL with a rearranged *MLL* gene (6) and 130 pediatric patients with different acute myeloid leukemia (AML) subtypes [76]. The ALL subtypes included T-ALL without or with rearrangement of the *MLL* gene and 6 different B-ALL subtypes, including those with a rearranged *MLL* gene, with chromosomal translocations involving BCR/ABL, E2A/PBX1 or TEL/AML1, with a hyperdiploid number of chromosomes (HD50) and others [81]. There were 6 different AML subtypes including those with a rearranged *MLL* gene, with chromosomal translocations involving PML/RARα, AML1/ETO or CBFβ/MYH11, M7 megakaryocytic leukemia and others [76].

Gene expression in these data sets was measured with the Affymetrix HG-U133A array. For all data sets, the expression value for each gene was determined using the MicroArray Suite version 5.0 (MAS 5.0) software [82].

Clustering of Highly Variable Genes in Normal Human Tissues

Expression values of PS in the duplicates of each normal tissue sample were averaged, expression values <20 were adjusted to 20 to eliminate noise from the data and all values were then $\log_{10}$ transformed. The 33,440 PS were filtered to select those genes that show a highly variable expression level in the 72 human tissue samples. We used two criteria to filter the PS and those PS that satisfied either criterion, were included: I. High (≥0.4) standard deviation of the log-transformed expression (LTE), measured over the different human tissues; II. LTE range of at least 2 *and* LTE value of ≥3 standard deviations below or above the mean in at least one tissue.

The 4,346 highly variable PS that passed either of these two criteria, were clustered according to their expression in the different human tissues, using the coupled two-way clustering (CTWC) algorithm [13]. Stable clusters were identified by CTWC after applying the Mean Field Approximation version [83] of the Super Paramagnetic Clustering (SPC) algorithm [12]. CTWC was applied using the default parameters, except for a minimal cluster size of 10. Prior to the first clustering step, the LTE of each gene was centered and normalized over the samples used.

Identification of Highly Expressed Genes

We used all the available PS to calculate separately the 85th percentile of the un-normalized expression values for each of the normal tissue samples and cancer cell lines. In the case of leukemic cells from patients, we first averaged the expression values over all patients with the same leukemia subtype, and took the 85th percentile of these average values. All the PS that were expressed at values higher than this threshold were defined as *highly expressed*. We have shown previously that changing the threshold used to define highly expressed genes to the 80th or 90th percentiles, did not affect the conclusions drawn from the analysis [79].

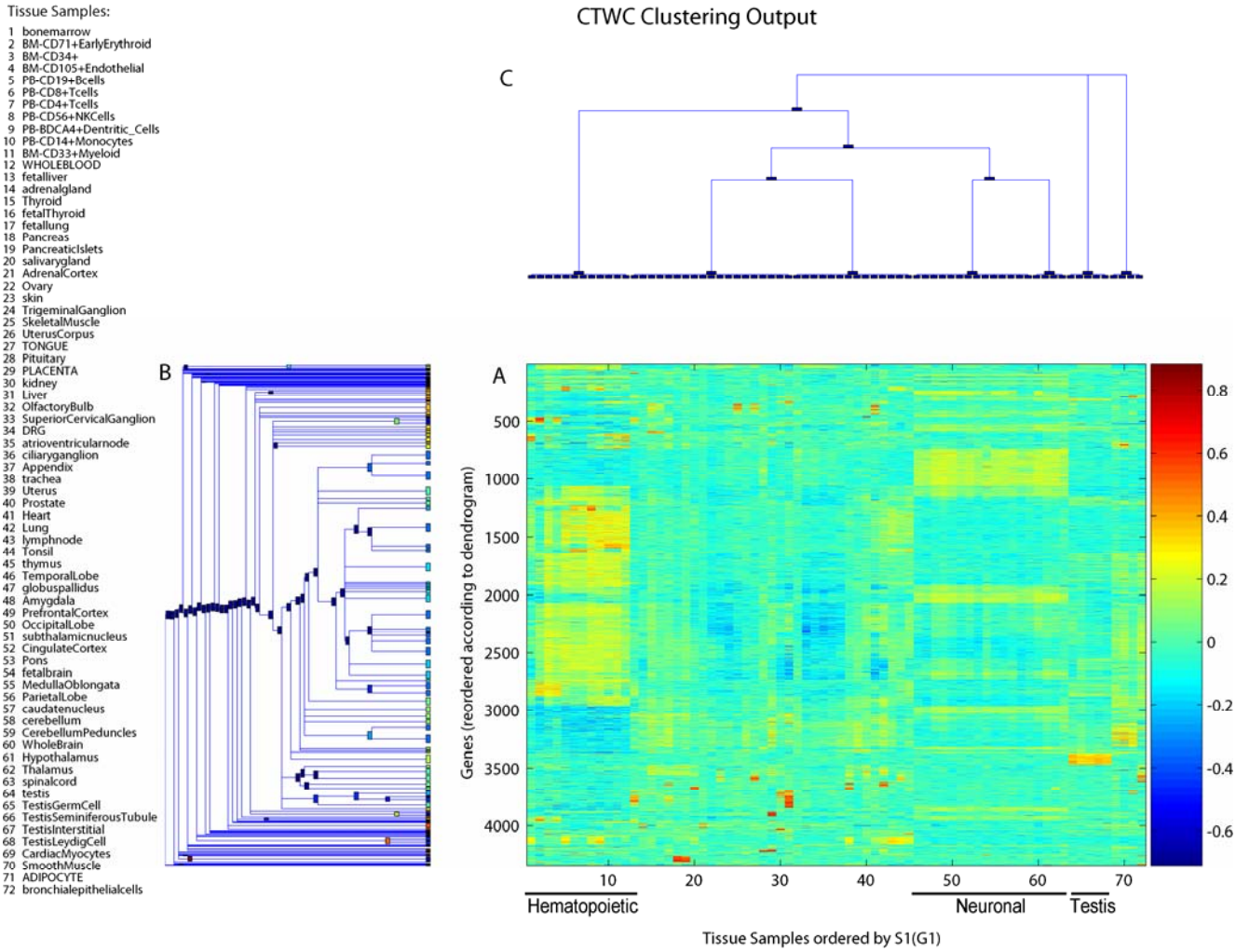


Figure 4. CTWC output for clustering highly variable genes and list of sample names. (A) Expression matrix ordered by genes (rows) and by samples (columns). (B) Gene clustering dendrogram. (C) Sample clustering dendrogram. The 72 sample types (expression matrix columns) are listed on the left.

Results

Clustering of Highly Variable Genes in Normal Human Tissues

After filtration, the 4,346 PS that showed highly variable expression in the 72 different normal tissue samples were clustered by CTWC. This clustering operation yielded 91 stable gene clusters that show *preferential expression* in different tissues (see Fig. 4). The term "preferential expression" refers to high relative expression levels of the majority of a cluster's genes in a particular subset of the samples, which was determined by inspection of color-coded expression matrices such as the left panels of Figs. 5 and 6. The genes included in 14 of these clusters showed preferential expression only in hematopoietic tissues. Further sub-classification within the hematopoietic tissues indicated that some of these clusters contained genes preferentially expressed either in T cells (see left panel, Fig 5A), B cells (Fig. 5B), myelomonocytic cells (Fig 5C) or erythroid and bone marrow endothelial cells (Fig. 5D), whereas other clusters contained genes expressed at similar levels in most hematopoietic cell types (Fig. 5E). In addition, 10 other clusters contained genes that were preferentially expressed in hematopoietic tissues plus 1-2 other tissues. For further analysis we shall refer to these 24 clusters (10 hematopoietic tissues only and 14 hematopoietic plus 1-2 other tissues) as hematopoietic (H) clusters (see the list of H clusters in Appendix II, Table II-1).

In addition to the H clusters, there were 28 clusters that contained genes preferentially expressed in various non-hematopoietic (NH) tissues. Some of these NH clusters, contained genes preferentially expressed in only one type of tissue such as neuronal, testis, placenta (Fig. 6 A-C, respectively, left panels), kidney, adrenal, pancreas or thyroid. Other NH clusters contained genes showing preferential expression in 2-3 tissues such as neuronal and testis, or in $\geq$4 different non-hematopoietic tissues (Fig. 6 D, left panel and table II-2 in appendix II).

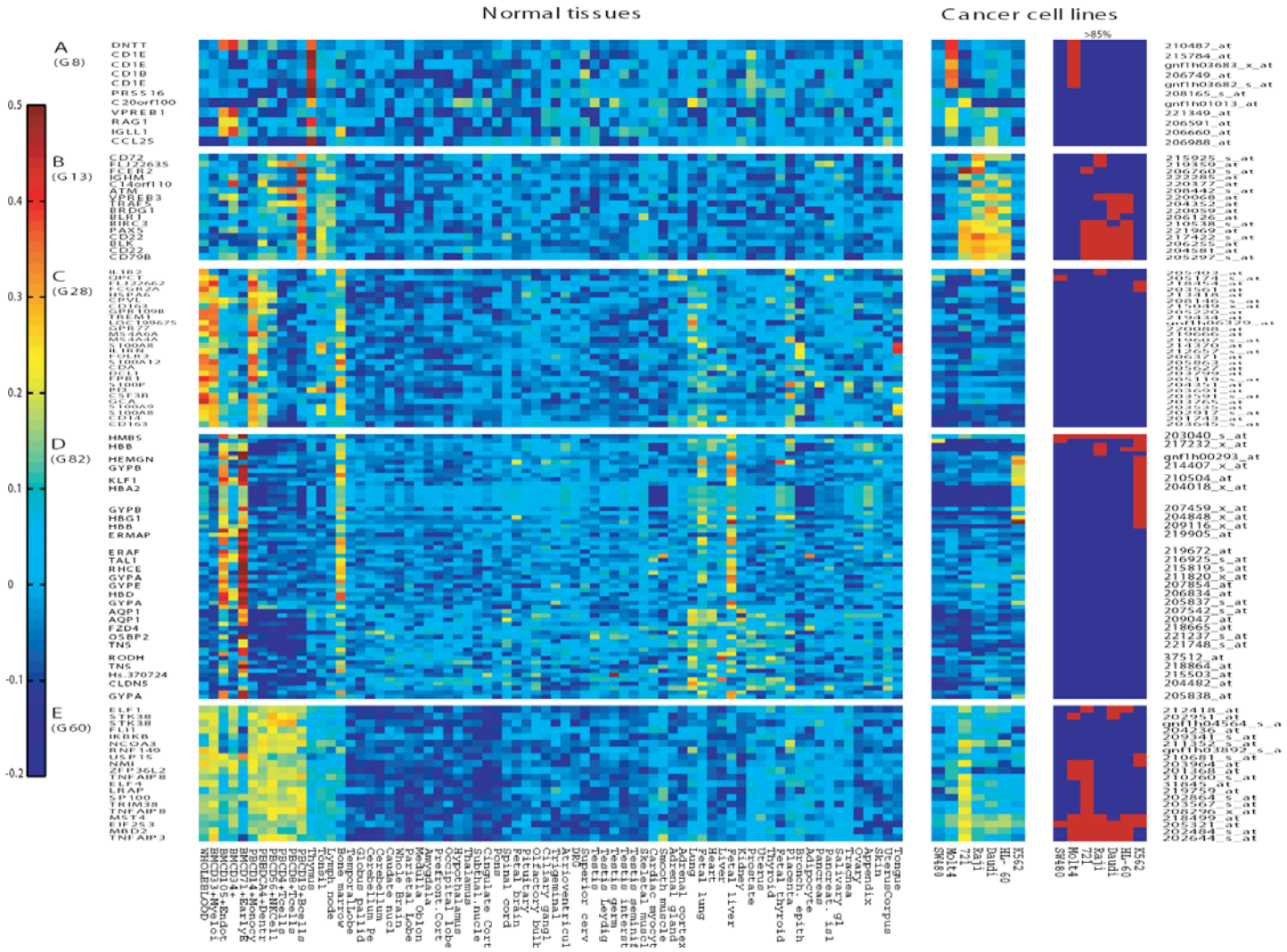


Figure 5. Examples of different H (Hematopoietic) clusters. Genes that show highly variable expression level in 72 normal tissue samples were clustered by CTWC as described in *Materials and Methods*. Some clusters showing preferential expression in normal hematopoietic tissues and the relative gene expression levels in seven human cancer cell lines are shown in *Left* and *Middle*, respectively, according to the color code shown on the left. Highly expressed genes in the cancer cell lines (>85%) are shown in *Right*, marked as red boxes. Because of the large size of some clusters, not all the gene names and their PS are marked on the left and right sides of the colored expression matrices, respectively.

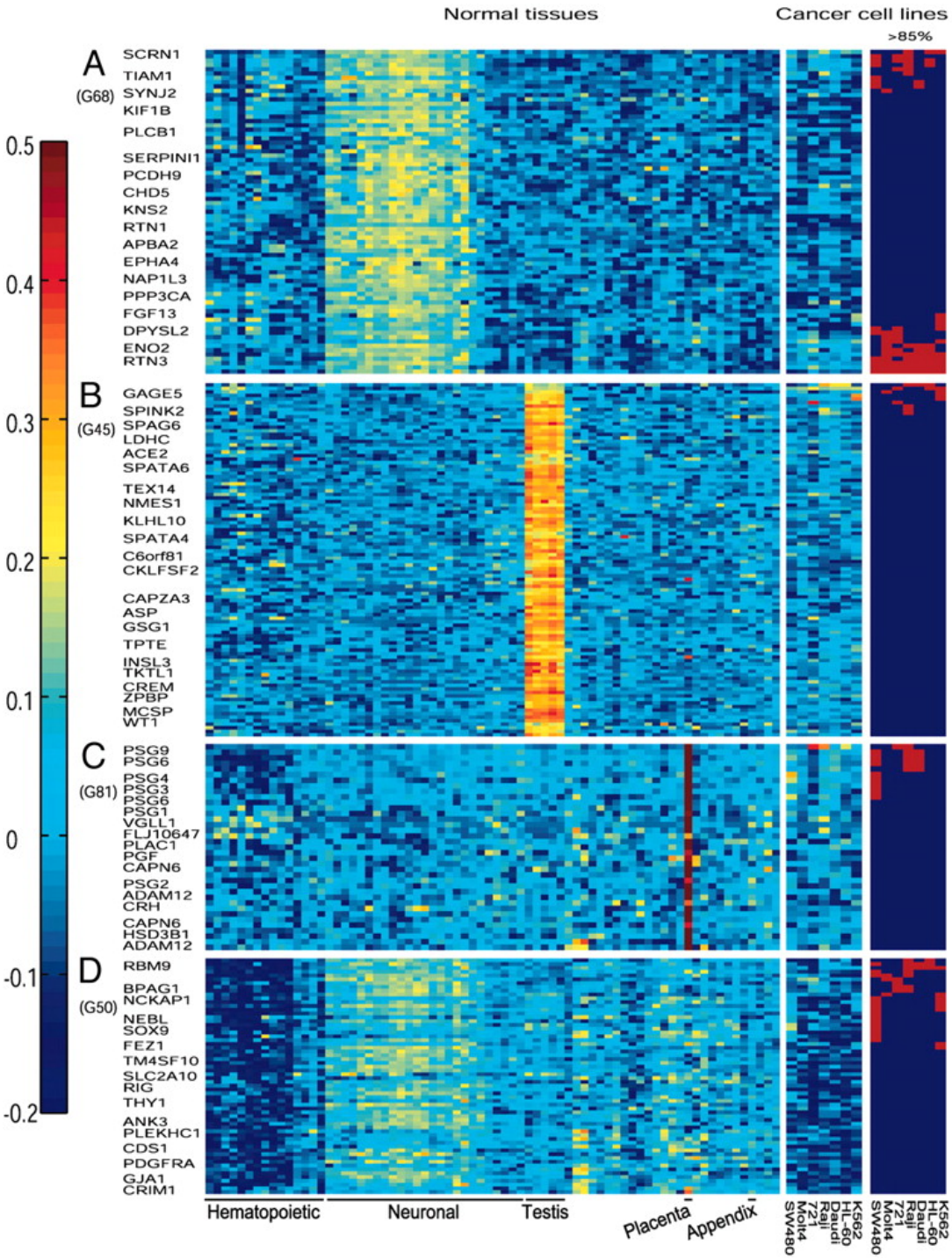


Figure 6. **Examples of different NH (Non-Hematopoietic) clusters.** Some clusters showing preferential expression in various normal non-hematopoietic tissues and the relative gene expression levels in seven human cancer cell lines are shown in *Left* and *Middle*, respectively, according to the color code shown on the left. Highly expressed genes in the cancer cell lines (>85%) are shown in *Right*, marked as red boxes. As in Fig. 5, not all the gene names are marked on the left side of the colored expression matrices. The order of the normal tissue samples and cancer cell lines is as in Fig. 5.

Testing for Distortion due to Normalization

The clustering operation that identified the H and NH clusters was based on Log-transformed expression (LTE) values that were centered and normalized, for each PS. This step may distort the relative expression levels of the genes in a particular sample. To show that this is not the case, we checked the overlap of the H and NH cluster genes with those that are identified as highly expressed genes, applying our standard 85th percentile threshold on the raw LTE values. The results indicate that 92.5% of the H cluster genes, i.e. 1046 out of 1130, were highly expressed in some hematopoietic tissues (Table 1). In contrast, only 3.5% of the H-cluster genes, 40 genes, were highly expressed in all normal tissues. There was also a low frequency of H-cluster genes that were highly expressed in various non-hematopoietic tissues, and for example, there were 126 such highly expressed genes in appendix (Table 1). An illustration of this phenomenon in two H clusters is shown in the Fig. 7 A and B.

Similar to the H clusters, 95.3% of the NH cluster genes, 1533 out of 1609, were highly expressed in the corresponding non-hematopoietic tissues, but only 1.8% of the NH cluster genes, 29 genes, were highly expressed in all normal tissues and only 273 of the NH cluster genes, were highly expressed in hematopoietic tissues (Table 1). An illustration of this phenomenon in two NH clusters is shown in Fig. 7 C and D. These results indicate that the genes included in the various H or NH clusters according to their *normalized* expression profile in different tissues, were also highly expressed mainly in the corresponding hematopoietic and non-hematopoietic tissues (based on the 85th percentile threshold applied on the *non-normalized* data).

	Normal tissues				Cancer cell lines						
Clusters	n	H	NH	Appendix	Molt4	721	Raji	Daudi	HL-60	K562	SW480
Hem.	1,130	1,046	–	126	226 (0)	378 (7)	260 (4)	281 (3)	247 (2)	165 (3)	154 (107)
Non-hem.	1,609	273	1,533	295	54 (20)	70 (34)	70 (24)	53 (11)	46 (17)	110 (50)	162 (110)

Table 1. Number of highly expressed PS in normal tissues and different human cancer cell lines. *n*, total number of PS in all H or NH clusters. Values in parentheses are the number of PS that were highly expressed in cancer cell lines but not in their normal counterparts.

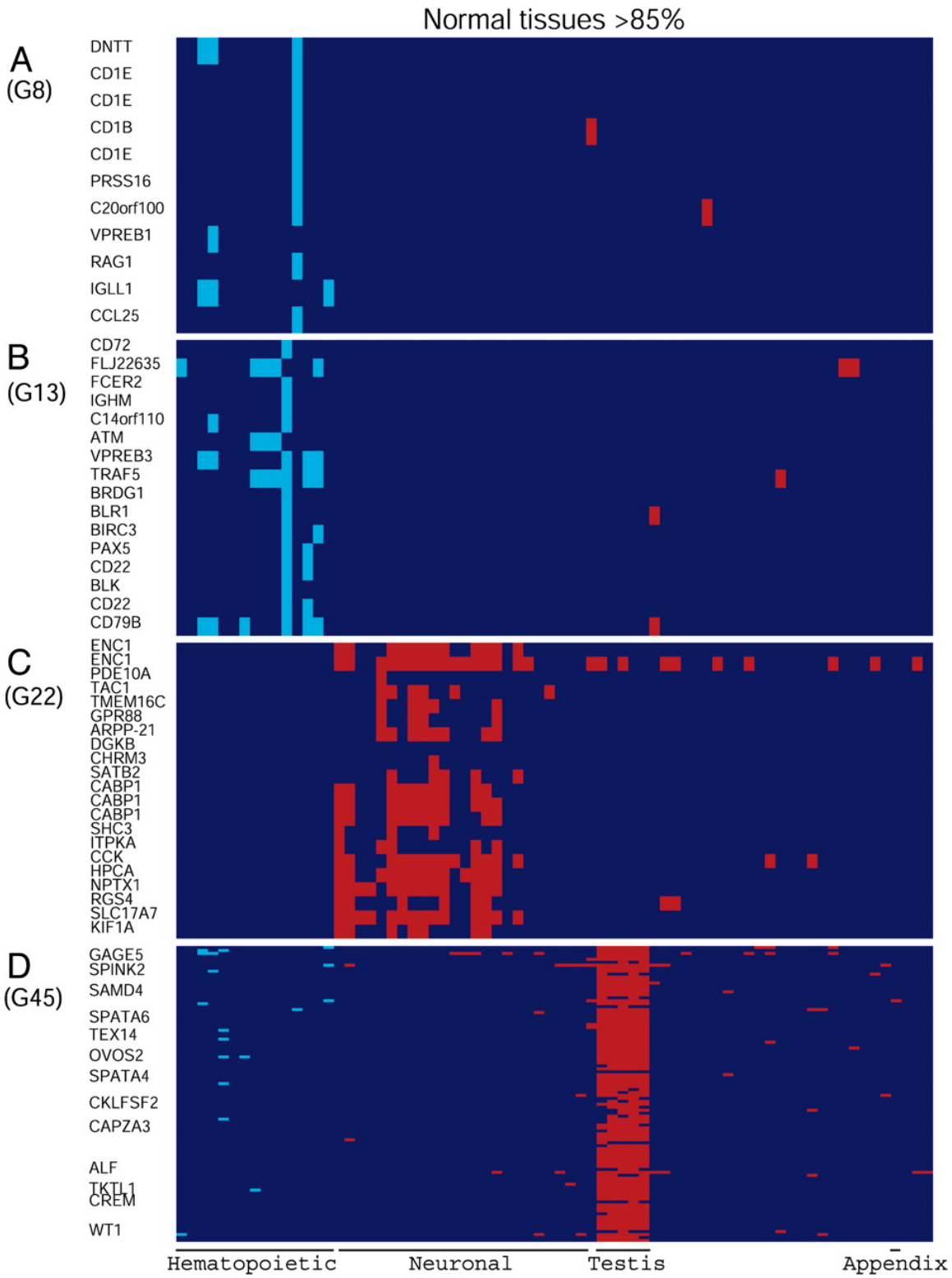


Figure 7. Highly expressed genes in some H and NH clusters in normal tissues. (*A* and *B*) H clusters. (*C* and *D*) NH clusters. Genes that are highly expressed in normal hematopoietic tissue samples or in any of the other normal tissues are marked as pale blue or red boxes, respectively. As in figure 5, not all the gene names are marked on the left side of *D*. The order of the normal tissue samples is as in figure 5.

Search for Genes that are highly expressed in Leukemic Cell Lines and in Some Normal Tissues, but *not* in Normal Hematopoietic Tissues

To compare gene expression in the normal tissues and cancer cell lines, we standardized the expression values of each gene over all the normal tissues and the 7 cancer cell lines included in the data set [80] (Figs. 5 and 6, left and middle panels, respectively). In these Figures, the internal ordering of the genes within each H and NH cluster was based on hierarchical clustering applied over the cancer cell lines. Thus, within each H or NH cluster, genes with similar expression profile over the cancer cell lines are adjacently placed (for ease of inspection). On the right panels of Figs. 5 and 6 we mark the genes that were highly expressed in the cancer cell lines.

The results with H-cluster genes indicate that different leukemia/lymphoma cell lines varied in the number of highly expressed genes (Table 1). As expected from their hematopoietic origin, almost all the genes that were highly expressed in the different cell lines were also highly expressed in their normal hematopoietic counterparts (Table 1). Those few H-cluster genes that were highly expressed in the leukemia/lymphoma cell lines but not in any of the normal hematopoietic counterparts are listed in Table II-3 on Appendix II. The results also show that only the T cell leukemia Molt4 highly expressed some genes that are preferentially expressed in normal thymus (Fig. 5 A). Furthermore, only B cell lymphomas 721, Raji and Daudi and the myeloid leukemia HL-60 highly expressed some genes that are preferentially expressed in normal B cells (Figs. 5B). In addition, the K562 cell line that can be induced to differentiate along the erythroid lineage [84], showed the highest number of highly expressed genes that are preferentially expressed in normal erythroid precursor cells (Fig. 5D). These results indicate that the human leukemia cell lines maintain certain features that characterize their normal cell lineage.

The behavior described above is the standard and expected one; high expression of genes in cancer cell lines is accompanied by high expression in the normal

tissue of origin. We turned to search for a more interesting, non-standard expression pattern, that of genes that are highly expressed in some cancer cell lines, low in the normal tissue from which these cancers originated, but high in some other normal tissue. Here we demonstrate that this non-standard scenario is indeed observed, concentrating first on leukemia/lymphoma cell lines. First, we noted that the NH-cluster genes exhibit variability in their expression over the different leukemia/lymphoma cell lines. For each of these different cell lines we identified the highly expressed genes that belong to the NH clusters. The leukemia/lymphoma cell lines highly expressed some NH cluster genes that are preferentially expressed in normal neuronal, testis, placenta (Fig. 6 A-C respectively), liver, kidney, thyroid, lung and some others. The different leukemia/lymphoma cell lines varied in the number of the highly expressed NH cluster genes (Table 1). For each cell line, we calculated the fraction of its highly expressed NH genes that were not highly expressed in any of the normal hematopoietic tissue samples. This fraction represents genes that are over expressed in the cell lines and it varied from 20% to 48% in different cell lines (Table 1 and see the list of genes in Table II-4, appendix II). Note that of the NH-cluster genes that were over-expressed in the cell lines, almost 80% were over expressed only in a single leukemic cell line (Table II-4). These results indicate that different human leukemia/lymphoma cell lines over-express genes that are normally preferentially expressed in tissues other than the hematopoietic tissue from which the leukemias originated.

Identification of Genes that are Over Expressed in Leukemic Cells from Human Patients with Different Subtypes of Lymphoid or Myeloid Leukemia

Cancer cells from patients with various types of cancer, including leukemia, over express various genes compared to their normal tissue of origin [73, 75-78, 81]. In view of our previous results [79] and the results described above with leukemia/lymphoma cell lines, we have now determined the extent to which leukemic cells from patients with different leukemia subtypes also highly express genes that are normally preferentially expressed in non-hematopoietic tissues. The gene expression data used in normal human tissues [80], leukemic cells from pediatric patients with ALL [81] or AML [76] were from different data sets, which prohibited direct comparison of gene expression values in these different studies. Therefore, we first calculated the average expression values of every gene in the data sets from all patients with the same leukemia subtype, and identified the highly expressed genes. We then determined which of the H or NH-cluster genes were highly expressed in the different leukemia subtypes.

Of the 1130 PS included in the normal H clusters, 1038 PS were also present in the human ALL and AML data sets. Of these 1038 H-cluster PS, the number of those that were highly expressed in leukemic cells from the different leukemia subtypes varied (over the subtypes) between 244 and 329, and almost all of these PS were also highly expressed in normal hematopoietic tissues (Table 2).

A

		T-ALL		B-ALL					
Clusters	**n**	**+MLL**	**–MLL**	**+MLL**	**BCR/ABL**	**E2A/PBX1**	**TEL/AML1**	**HD50**	**Others**
Hematopoietic	1,038	286 (0)	286 (1)	265 (0)	302 (0)	249 (1)	300 (0)	278 (0)	294 (0)
Nonhematopoietic	1,450	51 (10)	33 (3)	50 (9)	51 (5)	53 (9)	49 (13)	45 (6)	44 (4)

B

		AML					
Clusters	**n**	**+MLL**	**PML/RARα**	**AML1/ETO**	**CBFβ/MYH11**	**M7**	**Others**
Hematopoietic	1,038	318 (1)	244 (0)	277 (0)	329 (0)	281 (0)	306 (0)
Nonhematopoietic	1,450	44 (4)	53 (6)	41 (1)	35 (2)	55 (5)	30 (1)

Table 2. Number of highly expressed PS in different subtypes of human ALL (A) and AML (B). *n*, total number of PS in all H or NH clusters. Values in parentheses are the number of PS that are highly expressed in cells from different leukemia subtypes but not in normal hematopoietic cells.

Turning again to search for genes with non-standard expression patterns, we analyzed the NH-cluster genes, and found that the different ALL and AML subtypes highly expressed 30-55 genes, which were preferentially expressed in various normal non-hematopoietic tissues (Table 2). Of all the NH-cluster genes that were highly expressed in the different leukemias, 42 genes were not highly expressed in any normal hematopoietic tissue and are thus over expressed in the leukemias (Table 2 and see the list of these genes in Table II-5, appendix II). Of these 42 over expressed NH-cluster genes, 30 genes were over expressed only in a single leukemia subtype, and only 4 genes in ≥4 leukemia subtypes (Table II-5). These results indicate that like the leukemia/lymphoma cell lines, leukemic cells from patients also over express various genes that are normally preferentially expressed in various non-hematopoietic tissues including neuronal, testis, liver and placenta, and most of these genes were over expressed in just a single leukemia subtype.

Identification of Genes that are over Expressed in SW480 Adenocarcinoma Cell Line

The ability of leukemic cells to over express genes that are normally preferentially expressed in various non-hematopoietic tissues raised the question whether other types of cancer cells also possess this property. The results indicate that SW480 cells highly expressed various H-cluster genes (Table 1 and Fig. 5 C, D and E), although they had a lower number of such genes compared to the leukemia/lymphoma cell lines (Table 1). Furthermore, 69% of the highly expressed H-cluster genes in SW480 cells, were not highly expressed in the appendix (Table 1), which we used as their normal counterpart, and are thus over expressed in SW480. The H cluster genes that are over expressed in SW480 (Table II-3) include the apoptosis inhibitors *SERPINA1* and *BIRC5* and some genes involved in human cancer-associated translocations such as *LMO2*, *RUNX1* and *TCF3* that could play a role in their cancer phenotype.

Analysis of genes that are included in various NH clusters showed that SW480 highly expressed more such genes than the leukemic cell lines (Table 1). However, unlike the leukemia/lymphoma cell lines, SW480 did not highly express any of the genes that are preferentially expressed in normal testis (Fig. 6B and Table II-4). Of the 162 NH cluster genes that were highly expressed in SW480, 68% were not highly expressed in normal appendix (Table 1). In addition, only 21 of these over expressed NH cluster genes were common to SW480 and at least one of the leukemia or lymphoma cell lines (Table II-4). The results indicate that human adenocarcinoma cells, like leukemia/lymphoma cells, over-express many genes that are normally preferentially expressed in tissues other than their tissue of origin. As with the H cluster genes, the list of the over expressed NH cluster genes in SW480 includes many genes that are known to be over expressed in various types of human cancer and could contribute to cancer development and progression including *HOXA9, HOXB6, SOX9, CCND1, EGFR, SERPINE1, KRT8, KRT18, KRT19, TIAM1, FHL2* and *L1CAM* (Table II-4).

Discussion

Normal hematopoietic stem cells express genes that are preferentially expressed in various normal non-hematopoietic tissues [41]. We have previously shown that cells of a mouse myeloid leukemia cell line also highly express genes that are normally preferentially expressed in non-hematopoietic tissues such as neuronal, testis, liver and muscle tissues [79]. It is well established that human cancer cells from patients over express various genes compared to their normal tissue of origin [73, 75-78, 81]. We have determined now to what extent do different types of human cancer cells over express genes that are normally preferentially expressed in hematopoietic and non-hematopoietic tissues. We clustered genes that showed a highly variable expression level in 72 different normal human tissue samples and selected 2 major cluster categories; H clusters with genes preferentially expressed only in hematopoietic tissues or in hematopoietic tissues plus 1-2 other tissues, and NH clusters with genes preferentially expressed in a single or multiple non-hematopoietic tissues. More than 92% of the genes included in all H or NH clusters were highly expressed in the corresponding normal tissues based on the 85^{th} percentile criterion we have defined.

We determined which of the H or NH cluster genes were highly expressed in each of the human cancer cell lines tested and in different human ALL and AML subtypes. The results with H cluster genes indicated that different leukemia/lymphoma cell lines and leukemic cells from ALL and AML patients showed good lineage fidelity. As expected from the large fraction of H cluster genes that are highly expressed in normal hematopoietic tissues, almost all the H cluster genes that were highly expressed in the leukemia/lymphoma cell lines and leukemia patients were also highly expressed in normal hematopoietic cells. The colon adenocarcinoma SW480 also highly expressed 154 H cluster genes, but 69% of these genes were not highly expressed in normal appendix, which we used as a normal counterpart of SW480, and are thus over expressed in

SW480 compared to normal appendix. Two of these H cluster genes that are over expressed in SW480 cells, *SERPINA1*/*α1 ANTITRYPSIN* and *BIRC5*/*SURVIVIN* are anti-apoptotic genes [85, 86]. Furthermore, *SURVIVIN* is over expressed in many human cancers [86] including colorectal cancer in which it is regulated by the TCF/β-catenin pathway [87] and contributes to the radiation resistance in SW480 cells [88]. Some of the other H cluster genes that are over expressed in SW480 such as *LMO2*, *RUNX1/AML1* and *TCF3/E2A* are involved in human cancer-associated translocations [89-91]. Other H cluster genes that are over expressed in SW480, are also over expressed in a variety of human cancers and could contribute to development and progression of cancer due to their functions in regulating cell viability, proliferation, DNA repair, adhesion and invasiveness (Table II-6 on appendix II).

The results with NH cluster genes indicate that the leukemic cell lines, the ALL and AML leukemia subtypes and SW480 highly expressed various genes that are preferentially expressed in tissues other than those from which the cancers originated including neuronal, liver, kidney, thyroid, lung or placenta. The results have also indicated that a large proportion of the NH cluster genes that are highly expressed in the different cancer cells were over expressed in the cancer cells compared to their tissue of origin. Most of these genes were over expressed only in a single cancer cell line or leukemia subtype, indicating that the different cancer cells show differences both in the number and the identity of their over expressed genes. Many of these genes are up regulated in various types of human cancer and could contribute to cancer development and progression (See appendix II, Tables II-4 and I-7). In addition, it was reported by others that some of these genes including *SOX9, CCND1, EGFR, SERPINE1, TIAM1, FHL2* and *L1CAM* are indeed highly expressed in SW480 cells [64, 92-97]. Furthermore, *CCND1* and *L1CAM* are targets of the TCF/β-catenin pathway [64, 97], which is aberrantly activated in various cancers including colorectal cancer from which SW480 cells were derived.

It is suggested that the ability to over express genes that are normally preferentially expressed in tissues other than the cancer's origin, is a general property of cancer cells that plays a major role in determining the behavior of the cancers, including their metastatic potential. The results from the pediatric ALL and AML patients indicate that the ALLs over expressed more NH cluster genes than the AMLs, including genes preferentially expressed in neuronal tissues and testis. It will be interesting to find out whether this phenomenon is associated with the higher frequency of leukemia involvement in the central nervous system in ALL versus AML pediatric patients [76, 98]. Furthermore, the fact that a given cancer over expresses genes that are characteristic of a different normal tissue (other than its tissue of origin), may imply the existence of a differentiation therapy path which may be useful for differentiation therapy. The cancer's malignancy may be reduced by inducing it to trans-differentiate into this other normal tissue.

In the present study we scored genes that are over expressed in cancer cells using a very stringent threshold requirement, namely, only those genes that are above the 85^{th} percentile in the cancer cells but below this threshold in their normal counterparts. Therefore, all the over expressed genes we scored in cancer cells are also highly expressed genes. It is expected that there are other over expressed genes in cancer cells, whose level of expression in both normal and cancer cells is either above or below the 85^{th} percentile. Our results also indicate that there were differences in the identity of most of the over expressed genes between different cancer cell lines, even between leukemic cell lines from the same lineage. Therefore, the fact that we used the average expression values of genes from all patients with a given leukemia subtype, presumably resulted in detection of only a fraction of over expressed genes, those that are commonly over expressed in many of the patients. It is expected that additional over expressed genes can be identified, which show patient to patient differences even with the same leukemia subtype.

References

1. Heller, M.J., *DNA MICROARRAY TECHNOLOGY: Devices, Systems, and Applications.* Annual Review of Biomedical Engineering, 2002. **4**(1): p. 129-153.
2. Affymetrix. *Affymetrix HG-U133Av2 Data Sheet*. 2004 February 2006 [cited; Available from: http://www.affymetrix.com/support/technical/datasheets/human_datasheet.pdf.
3. Lipshutz, R.J., et al., *High density synthetic oligonucleotide arrays.* Nat Genet, 1999.
4. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus.* Nat Rev Genet, 2006. **7**(1): p. 55-65.
5. Huber, W., et al., *Variance stabilization applied to microarray data calibration and to the quantification of differential expression.* Bioinformatics, 2002. **18**(suppl_1): p. S96-104.
6. Speed, T. *Terry Speed's Microarray Data Analysis Group.* 2000 [cited; Available from: http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html.
7. Lowry, R., *Concepts and Applications of Inferential Statistics*. 2006.
8. StatSoft, I., *The Statistics Homepage*.
9. Benjamini, Y.a.H., Y, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society, 1995. **57**: p. 289-300.
10. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nat Genet, 2000. **25**(1): p. 25-29.
11. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
12. Blatt, M., S. Wiseman, and E. Domany, *Superparamagnetic clustering of data.* Physical Review Letters, 1996. **76**(18): p. 3251-3254.
13. Getz, G., E. Levine, and E. Domany, *Coupled two-way clustering analysis of gene microarray data.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12079-84.
14. Getz, G., et al., *Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data.* Bioinformatics, 2003. **19**(9): p. 1079-89.
15. Getz, G. and E. Domany, *Coupled two-way clustering server.* Bioinformatics, 2003. **19**(9): p. 1153-4.
16. Malcolm R. Alison, R.P.S.F.N.A.W., *An introduction to stem cells.* The Journal of Pathology, 2002. **197**(4): p. 419-423.
17. Rossant, J., *Stem Cells from the Mammalian Blastocyst.* Stem Cells, 2001. **19**(6): p. 477-482.
18. Evans, M.J. and M.H. Kaufman, *Establishment in culture of pluripotential cells from mouse embryos.* Nature, 1981. **292**(5819): p. 154-156.
19. Thomson, J.A., et al., *Embryonic Stem Cell Lines Derived from Human Blastocysts.* Science, 1998. **282**(5391): p. 1145-1147.
20. Smith, A.G., *EMBRYO-DERIVED STEM CELLS: Of Mice and Men.* Annual Review of Cell and Developmental Biology, 2001. **17**(1): p. 435-462.
21. Pedersen, R.A., *Feeding hungry stem cells.* Nat Biotech, 2002. **20**(9): p. 882-883.
22. Brivanlou, A.H., et al., *STEM CELLS: Enhanced: Setting Standards for Human Embryonic Stem Cells.* Science, 2003. **300**(5621): p. 913-916.

23. Eckfeldt, C.E., E.M. Mendenhall, and C.M. Verfaillie, *The molecular repertoire of the 'almighty' stem cell.* Nat Rev Mol Cell Biol, 2005. **6**(9): p. 726-37.
24. Ulloa-Montoya F, V.C., Hu WS., *Culture systems for pluripotent stem cells.* J Biosci Bioeng, 2005. **100**(1): p. 12-27.
25. Rajasekhar, V.K. and M.C. Vemuri, *Molecular Insights into the Function, Fate, and Prospects of Stem Cells.* Stem Cells, 2005. **23**(8): p. 1212-1220.
26. Wagers, A.J. and I.L. Weissman, *Plasticity of adult stem cells.* Cell, 2004. **116**(5): p. 639-48.
27. Pittenger, M.F., et al., *Multilineage Potential of Adult Human Mesenchymal Stem Cells.* Science, 1999. **284**(5411): p. 143-147.
28. Kuznetsov, S.A., et al., *Circulating Skeletal Stem Cells.* J. Cell Biol., 2001. **153**(5): p. 1133-1140.
29. Lee, O.K., et al., *Isolation of multipotent mesenchymal stem cells from umbilical cord blood.* Blood, 2004. **103**(5): p. 1669-1675.
30. Zuk, P.A., et al., *Human Adipose Tissue Is a Source of Multipotent Stem Cells.* Mol. Biol. Cell, 2002. **13**(12): p. 4279-4295.
31. Friedenstein, A.J., J.F. Gorskaja, and N.N. Kulagina, *Fibroblast precursors in normal and irradiated mouse hematopoietic organs.* Exp Hematol, 1976. **4**(5): p. 267-74.
32. Alhadlaq, A. and J.J. Mao, *Mesenchymal Stem Cells: Isolation and Therapeutics.* Stem Cells and Development, 2004. **13**(4): p. 436-448.
33. Lerou, P.H. and G.Q. Daley, *Therapeutic potential of embryonic stem cells.* Blood Reviews, 2005. **19**(6): p. 321-331.
34. Gafni, Y., et al., *Stem cells as vehicles for orthopedic gene therapy.* Gene Ther, 2004. **11**(4): p. 417-26.
35. Caplan, A.I., *Review: Mesenchymal Stem Cells: Cell–Based Reconstructive Therapy in Orthopedics.* Tissue Engineering, 2005. **11**(7-8): p. 1198-1211.
36. Ivanova, N.B., et al., *A stem cell molecular signature.* Science, 2002. **298**(5593): p. 601-4.
37. Ramalho-Santos, M., et al., *"Stemness": transcriptional profiling of embryonic and adult stem cells.* Science, 2002. **298**(5593): p. 597-600.
38. Fortunel, N.O., et al., *Comment on " 'Stemness': Transcriptional Profiling of Embryonic and Adult Stem Cells" and "A Stem Cell Molecular Signature" (I).* Science, 2003. **302**(5644): p. 393b-.
39. Evsikov, A.V. and D. Solter, *Comment on " 'Stemness': Transcriptional Profiling of Embryonic and Adult Stem Cells" and "A Stem Cell Molecular Signature" (II).* Science, 2003. **302**(5644): p. 393c-.
40. Golan-Mashiach, M., et al., *Design principle of gene expression used by human stem cells: implication for pluripotency.* Faseb J, 2005. **19**(1): p. 147-9.
41. 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. 383-389.
42. Arney, K.L. and A.G. Fisher, *Epigenetic aspects of differentiation.* J Cell Sci, 2004. **117**(19): p. 4355-4363.

43. Gerecht-Nir, S., et al., *Vascular gene expression and phenotypic correlation during differentiation of human embryonic stem cells.* Dev Dyn, 2005. **232**(2): p. 487-97.
44. Gerecht-Nir, S., et al., *Human embryonic stem cells as an in vitro model for human vascular development and the induction of vascular differentiation.* Lab Invest, 2003. **83**(12): p. 1811-20.
45. *Stem Cells: Scientific Progress and Future Research Directions*. 2001 [cited; Available from: http://stemcells.nih.gov/info/scireport.
46. Sekiya, I., et al., *Adipogenic differentiation of human adult stem cells from bone marrow stroma (MSCs).* J Bone Miner Res, 2004. **19**(2): p. 256-64.
47. Nuttall, M.E. and J.M. Gimble, *Controlling the balance between osteoblastogenesis and adipogenesis and the consequent therapeutic implications.* Current Opinion in Pharmacology, 2004. **4**(3): p. 290-294.
48. Xing, L. and B.F. Boyce, *Regulation of apoptosis in osteoclasts and osteoblastic cells.* Biochemical and Biophysical Research Communications, 2005. **328**(3): p. 709-720.
49. Park, Y. and S.L. Gerson, *DNA REPAIR DEFECTS IN STEM CELL FUNCTION AND AGING.* Annual Review of Medicine, 2005. **56**(1): p. 495-508.
50. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
51. Esteller, M., *Epigenetics provides a new generation of oncogenes and tumour-suppressor genes.* Br J Cancer, 2006. **94**(2): p. 179-83.
52. Paul G. Corn, W.S.E.-D., *Derangement of growth and differentiation control in oncogenesis.* BioEssays, 2002. **24**(1): p. 83-90.
53. Clarke, C.A., et al., *Acute Myeloid Leukemia.* N Engl J Med, 2000. **342**(5): p. 358-359.
54. Smith, M., et al., *Adult acute myeloid leukaemia.* Critical Reviews in Oncology/Hematology, 2004. **50**(3): p. 197-222.
55. Tenen, D.G., *Disruption of differentiation in human cancer: AML shows the way.* Nat Rev Cancer, 2003. **3**(2): p. 89-101.
56. Gilliland, D.G. and M.S. Tallman, *Focus on acute leukemias.* Cancer Cell, 2002. **1**(5): p. 417-20.
57. de The, H. and M.K. Chelbi-Alix, *APL, a model disease for cancer therapies?* Oncogene, 2001. **20**(49): p. 7136-9.
58. Lengfelder, E., et al., *Treatment concepts of acute promyelocytic leukemia.* Critical Reviews in Oncology/Hematology, 2005. **56**(2): p. 261-274.
59. Rabinowitz, Z. and L. Sachs, *Reversion of properties in cells transformed by polyoma virus.* Nature, 1968. **220**(173): p. 1203-1206.
60. Lotem, J. and L. Sachs, *Epigenetics wins over genetics: induction of differentiation in tumor cells.* Seminars in Cancer Biology, 2002. **12**(5): p. 339-346.
61. Shachaf, C.M., et al., *MYC inactivation uncovers pluripotent differentiation and tumour dormancy in hepatocellular cancer.* Nature, 2004. **431**(7012): p. 1112-1117.

62. Shachaf, C.M. and D.W. Felsher, *Tumor Dormancy and MYC Inactivation: Pushing Cancer to the Brink of Normalcy.* Cancer Res, 2005. **65**(11): p. 4471-4474.
63. Spira, A.I. and M.A. Carducci, *Differentiation therapy.* Current Opinion in Pharmacology, 2003. **3**(4): p. 338-343.
64. Cao, T. and B.C. Heng, *Differentiation therapy of cancer. Potential advantages over conventional therapeutic approaches targeting death of cancer/tumor cells.* Medical Hypotheses, 2005. **65**(6): p. 1202-1203.
65. Sell, S., *Cancer Stem Cells and Differentiation Therapy.* Tumor Biology, 2006. **27**(2): p. 59-70.
66. Soltysova, A., V. Altanerova, and C. Altaner, *Cancer stem cells.* Neoplasma, 2005. **52**(6): p. 435-40.
67. Marx, J., *CANCER RESEARCH: Mutant Stem Cells May Seed Cancer.* Science, 2003. **301**(5638): p. 1308-1310.
68. Romano, G., *The role of adult stem cells in carcinogenesis.* Drug News Perspect, 2005. **18**(9): p. 555-9.
69. Al-Hajj, M. and M.F. Clarke, *Self-renewal and solid tumor stem cells.* Oncogene, 2004. **23**(43): p. 7274-82.
70. Bjerkvig, R., et al., *The origin of the cancer stem cell: current controversies and new insights.* Nat Rev Cancer, 2005. **5**(11): p. 899-904.
71. Richard Poulsom, M.R.A.S.J.F.N.A.W., *Adult stem cell plasticity.* The Journal of Pathology, 2002. **197**(4): p. 441-456.
72. Krause, D.S., *Plasticity of marrow-derived stem cells.* Gene Ther, 2002. **9**(11): p. 754-8.
73. Jarzab, B., et al., *Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications.* Cancer Res, 2005. **65**(4): p. 1587-1597.
74. Ma, X.-J., et al., *Gene expression profiles of human breast cancer progression.* PNAS, 2003. **100**(10): p. 5974-5979.
75. Okutsu, J., et al., *Prediction of chemosensitivity for patients with acute myeloid leukemia, according to expression levels of 28 genes selected by genome-wide complementary DNA microarray analysis.* Mol Cancer Ther, 2002. **1**(12): p. 1035-1042.
76. Ross, M.E., et al., *Gene expression profiling of pediatric acute myelogenous leukemia.* Blood, 2004. **104**(12): p. 3679-87.
77. Sotiriou, C., et al., *Molecular profiling of head and neck tumors.* Curr Opin Oncol, 2004. **16**(3): p. 211-214.
78. Hong, J.-H., et al., *TAZ, a Transcriptional Modulator of Mesenchymal Stem Cell Differentiation.* Science, 2005. **309**(5737): p. 1074-1078.
79. Lotem, J., et al., *Induction in myeloid leukemic cells of genes that are expressed in different normal tissues.* Proc Natl Acad Sci U S A, 2004. **101**(45): p. 16022-7.
80. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.
81. Ross, M.E., et al., *Classification of pediatric acute lymphoblastic leukemia by gene expression profiling.* Blood, 2003. **102**(8): p. 2951-9.
82. Hubbell, E., W.M. Liu, and R. Mei, *Robust estimators for expression analysis.* Bioinformatics, 2002. **18**(12): p. 1585-1592.

83. Barad, O., *M.Sc.* 2003, The Weizmann Institute of Science: Rehovot, Israel.
84. Gahmberg, C.G. and L.C. Andersson, *K562--a human leukemia cell line with erythroid features.* Semin Hematol, 1981. **18**(1): p. 72-77.
85. Van Molle, W., et al., *Alpha 1-acid glycoprotein and alpha 1-antitrypsin inhibit TNF-induced but not anti-Fas-induced apoptosis of hepatocytes in mice.* J Immunol, 1997. **159**(7): p. 3555-3564.
86. Altieri, D.C., *Survivin and apoptosis control.* Adv Cancer Res, 2003. **88**: p. 31-52.
87. Lee, R.H., et al., *Characterization and Expression Analysis of Mesenchymal Stem Cells from Human Bone Marrow and Adipose Tissue.* Cellular Physiology and Biochemistry, 2004. **14**(4-6): p. 311-324.
88. Rodel, F., et al., *Survivin as a radioresistance factor, and prognostic and therapeutic target for radiotherapy in rectal cancer.* Cancer Res, 2005. **65**(11): p. 4881-4887.
89. Rabbitts, T.H., et al., *Chromosomal translocations and leukaemia: a role for LMO2 in T cell acute leukaemia, in transcription and in erythropoiesis.* Leukemia, 1997. **11 Suppl 3**: p. 271-272.
90. Roumier, C., et al., *New mechanisms of AML1 gene alteration in hematological malignancies.* Leukemia, 2003. **17**(1): p. 9-16.
91. LeBrun, D.P., *E2A basic helix-loop-helix transcription factors in human leukemia.* Front Biosci, 2003. **8**: p. s206-s222.
92. Blache, P., et al., *SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the CDX2 and MUC2 genes.* J Cell Biol, 2004. **166**(1): p. 37-47.
93. Partik, G., et al., *Inhibition of epidermal-growth-factor-receptor-dependent signalling by tyrphostins A25 and AG1478 blocks growth and induces apoptosis in colorectal tumor cells in vitro.* J Cancer Res Clin Oncol, 1999. **125**(7): p. 379-388.
94. Schwarte-Waldhoff, I., et al., *DPC4/SMAD4 mediated tumor suppression of colon carcinoma cells is associated with reduced urokinase expression.* Oncogene, 1999. **18**(20): p. 3152-3158.
95. Liu, L., D.H. Wu, and Y.Q. Ding, *Tiam1 gene expression and its significance in colorectal carcinoma.* World J Gastroenterol, 2005. **11**(5): p. 705-707.
96. Chan, K.K., et al., *Protein-protein interaction of FHL2, a LIM domain protein preferentially expressed in human heart, with hCDC47.* J Cell Biochem, 2000. **76**(3): p. 499-508.
97. Gavert, N., et al., *L1, a novel target of beta-catenin signaling, transforms cells and is expressed at the invasive front of colon cancers.* J Cell Biol, 2005. **168**(4): p. 633-642.
98. Dusenbery, K.E., et al., *Extramedullary leukemia in children with newly diagnosed acute myeloid leukemia: a report from the Children's Cancer Group.* J Pediatr Hematol Oncol, 2003. **25**(10): p. 760-768.

Appendix I

The effect of standardization on the ESC and MSC dataset

The following figure presents the dataset after all-absent removal, thresholding to 1 and applying Log2 transformation, for all the ESC and Mesenchymal samples. The genes were ordered by the average expression values of the ESC.

The dramatic effect of the standardization can be seen in Fig I-2.

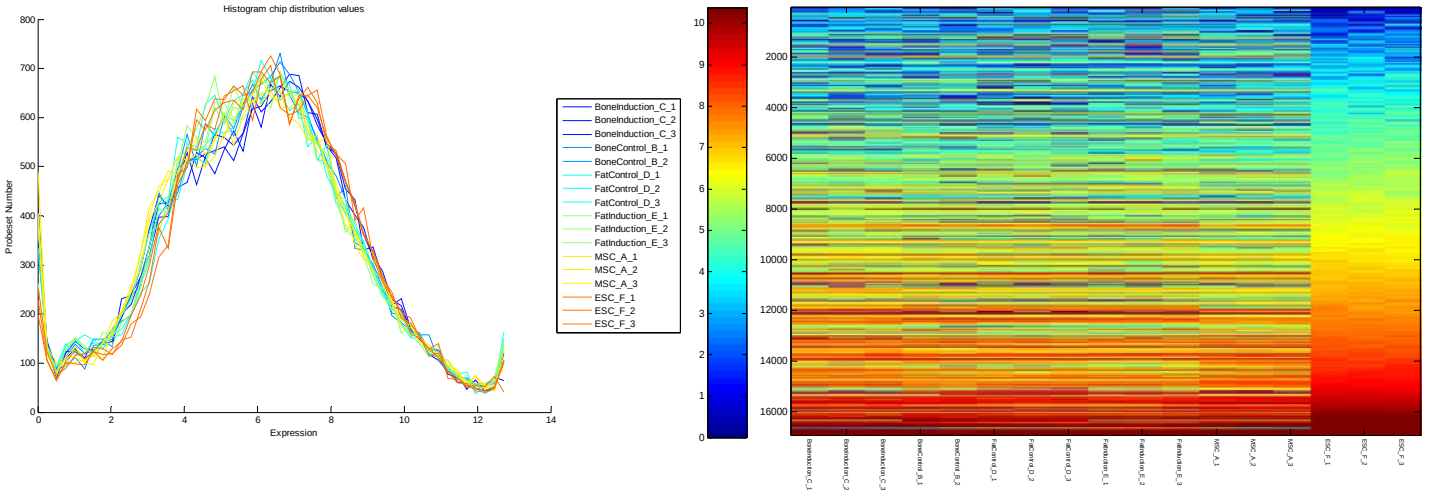


Figure I-1: Dataset histogram and expression matrix before standardization

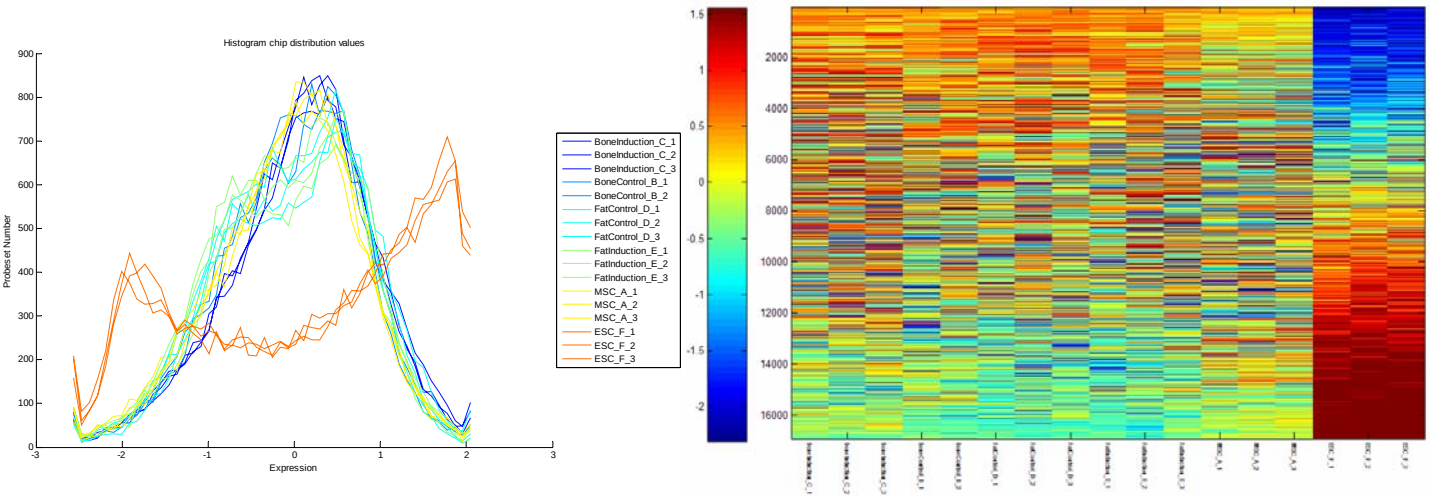


Figure I-2: Dataset histogram and expression matrix after standardize

The question that the above figures raise is as follows: If the ESC samples are so different from the mesenchymal samples (as can be seen on the standardized dataset histogram), how come the ESC samples look very similar in their distribution to the mesenchymal samples before standardization? For simplicity of presentation, we show from now on the histograms of the averaged expression values over the different replicates.

As mentioned in the general methods section, gene standardization is a two-step process: row centering followed by row normalization.

In centering, the average expression of the gene is substracted from each expression value, independently for every gene. In normalization, each expression value is divided by the gene's standard deviation, independently for every gene. That is, the standardized expression value E'_{gs} of the expression of gene g on sample s is given by

$$E'_{gs} = \frac{E_{gs} - \overline{E}_g}{\sigma_g} \quad \text{where} \quad \sigma_g = \sqrt{\frac{\sum_s (E_{gs} - \overline{E_g})^2}{n_s - 1}}$$

The following figures were created in order to try to understand how come the nearly similar distributions of the expression levels measured for the different kinds of cells (see Fig I-1) becomes, after standardization, so strikingly different for the ESCs. The effect of the standardization process is shown, one step at the time, in figure I-3, which represents the histograms of the expression values, averaged over each set of replicates.

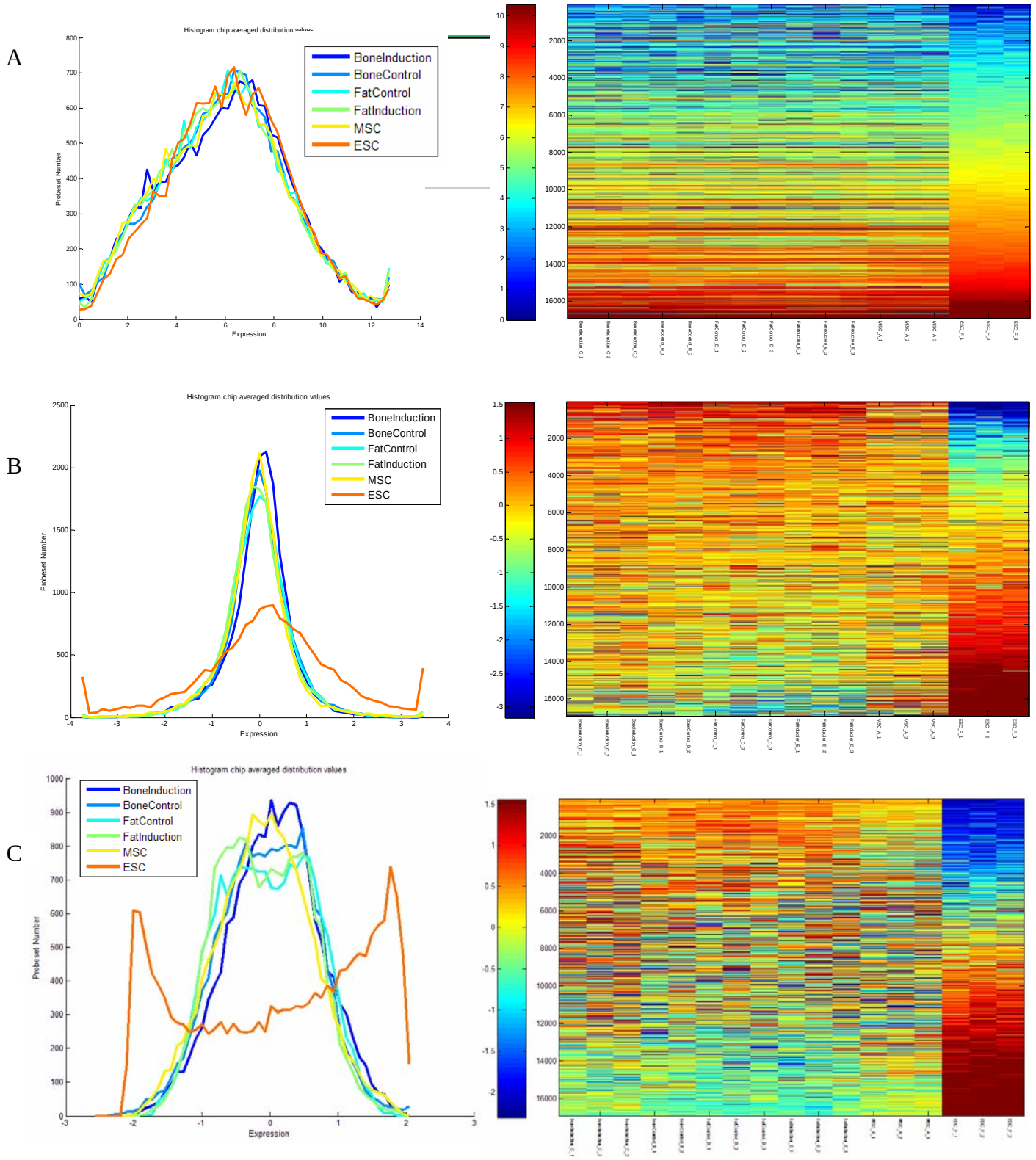


Figure I-3: Dataset expression matrix and corresponding distribution. Expression matrices on the right present expression levels for 16,461 dataset probe-sets in different preprocessing steps. Probesets are sorted by their averaged expression on the ESC samples. Histograms on the left show expression distribution (each curve represent replicate-averaged sample type). **(A)** Raw data after removal of 'all-absent' probesets, threshold of 1, log2 transformed. **(B)** Centered data **(C)** Standardized data (centering + normalization).

Comparing the expression histograms I-3A and I-3B one can see that this step alone distinguishes the ESC samples from the other, mesenchymal samples. The ESC distribution in I-3B is shifted slightly to values that are more positive and there is a much more prominent change: the distribution has much "fatter" tails on both sides of the peak, especially on the positive side. The high-end tail is caused by the fact that a large number of genes have higher expression levels in the ESCs than in the MSCs. There are also a (somewhat lower) number of genes for which the situation is the opposite; these generate the low-end tail.

Upon normalization, these fat tails are turn into the two peaks, at low and high normalized expression values, as see in Fig I-3C.

In order to gain further insight into the manner in which the expression levels and their distributions are modified by standardization, we generated the following set of figures, which show individual gene expression value for ESC, MSC, Fat-Control and Fat-Induction before and after standardization. Each dot represent a probe-set (averaged over replicates): red points represent probe-set expression values before standardization and blue dots represent expression values after standardization (transformed to fit in the same scale as the red points). Each sub-plot is sorted according to the pre-standardization values.

Comparing the four sub-plots of figure I-4, it is observed how the expression values of each sample type are transformed due to standardization. ESC expression values are strongly shifted up or down, MSC sample seem to be always in a “good” place in the middle (compared to other samples) and thus their post-standardization values are mainly concentrated on the middle. Standardization shifts the Fat sample expression values mildly up or down.

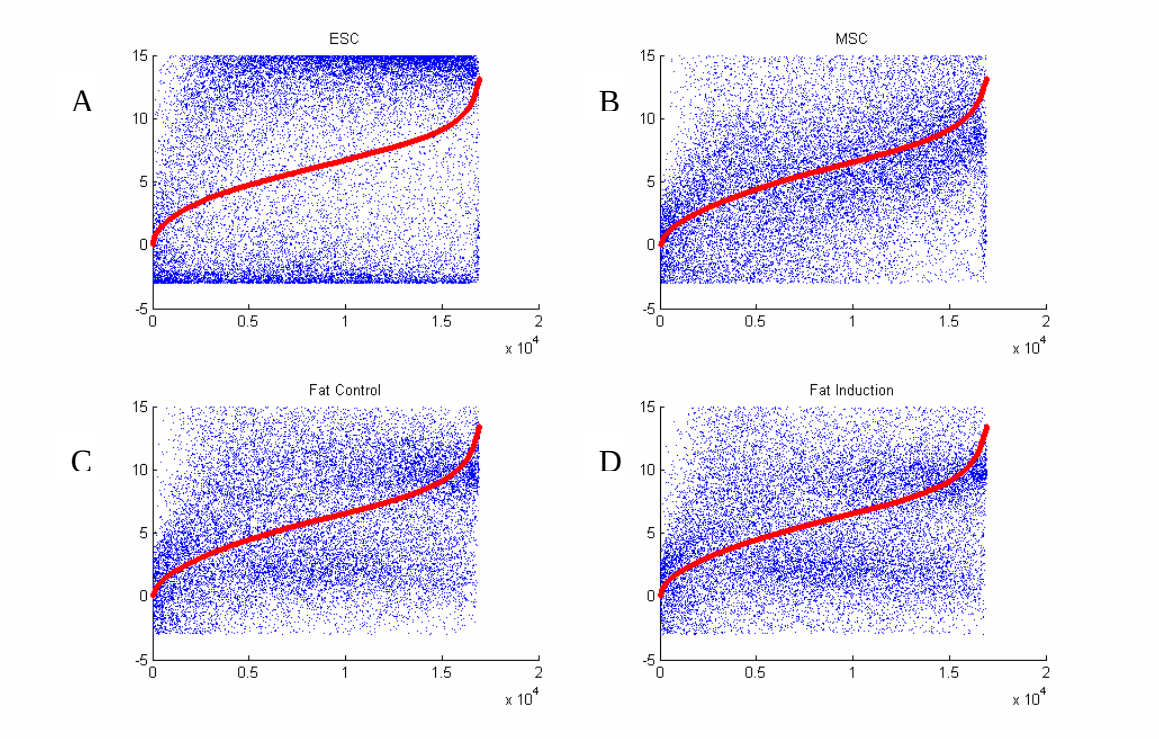


Figure I-4: the effect of standardization on the expression levels of individual genes, sorted according to their expression levels in each cell type (red) and after standardization (blue)

Next, to ascertain that the observed effect is indeed due to expression differences between ESCs and the other cells, we grouped the genes on the basis of their average un-standardized expression on the ESCs and studied the distribution of expression of each such group in the different mesenchymal cell types. The following series of figures displays the distributions of un-standardized expression levels of ESCs, MSCs, Fat-Control and Fat-Induction, divided to 5 groups of genes (bins), based on the ESC samples. The first figure shows all probe-sets whose ESC expression values is between 0 and 2.5, the second one corresponds to the range 2.5-5, and so forth.

We see clearly that the differences between ESCs and other samples are indeed found on the un-standardized data as well. Looking at the figure I-5D (ESC bin range of 7.5-10) as an example, we see that hundreds of genes are expressed significantly lower on MSC and Fat compared to ESC

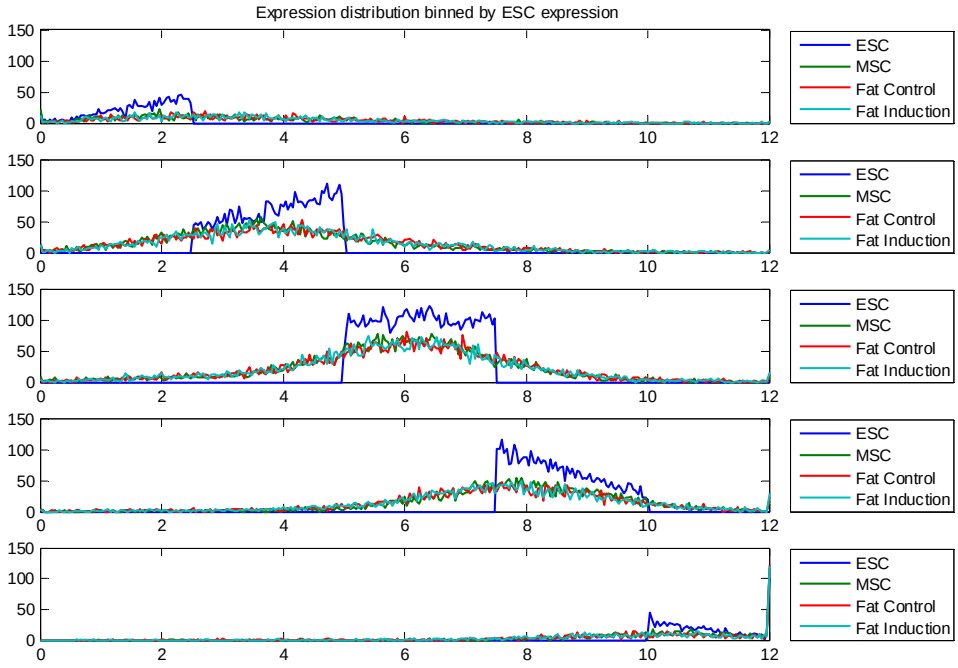


Figure I-5: Distributions of un-standardized expression values of different cell types. Genes were binned according to ESC expression levels.

The following two figures were created in order to test the hypothesis that the effect we saw (of standardization turning the distribution of expression on ESC samples into bi-moda), is caused by the fact that we have a relatively small number of ESC samples (3) versus the large number of mesenchymal samples (14), which are fairly similar to each other.

In the following graphs (I-6 and I-7), standardization was conducted on different subsets of the dataset samples in order to test whether the post-standardization distribution is affected by the number of samples in the dataset.

This hypothesis is rejected because it was observed that bi-modality is not dependent on the number of mesenchymal samples against which the ESC samples are standardized. Furthermore, bi-modality is not special for ESC samples, and this distribution is generated also by standardizing MSC samples together with Fat-Induction samples or others.

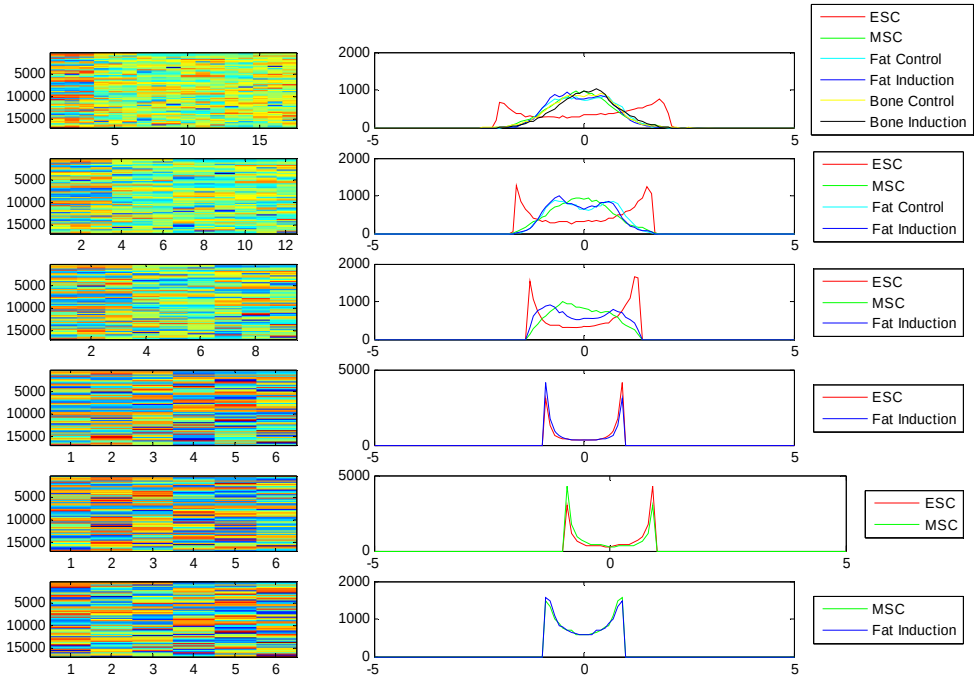


Figure I-6. Standardization of different sub-sets of the dataset samples – Fat samples.

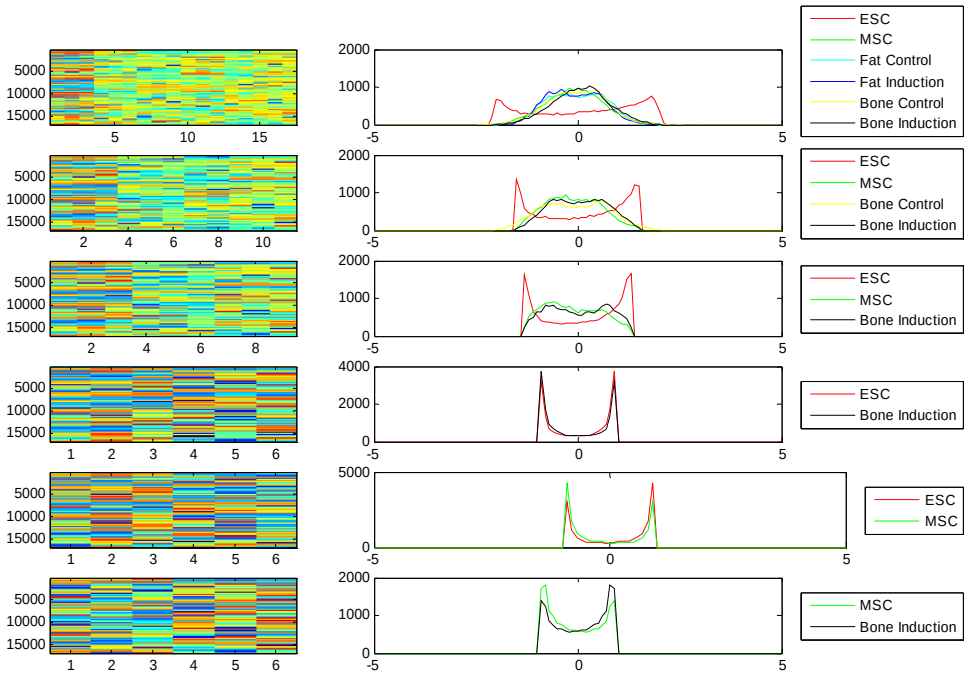


Figure I-7. Standardization of different sub-sets of the dataset samples – Bone samples.

Based on these and other graphs (not shown), we have concluded that the standardization transformation will take the most outlying expression values of each gene and move it further to the extreme. By an accumulative effect, samples whose expression levels are found to be most "coherently different" on many probe-sets, will pile up on the histogram edges, thus creating a bi-modal distribution.

Appendix II

Table II-1. Clusters of normal human tissues - Hematopoietic (H) clusters

No.	Cluster	No. of genes	Preferentially expressed in	Symbol
1	G8	11	Hemopoietic only (esp. in Thymus)	**H**
2	G13	16	Hemopoietic only (esp. in B cells)	**H**
3	G14	22	Hemopoietic only (esp. in B cells, DC, LN, Tonsil)	**H**
4	G28	28	Hemopoietic (esp. in BM, Mono, Myeloid, FL)	**H**
5	G30	42	Hemopoietic only (esp. in BM, WB)	**H**
6	G60	20	Hemopoietic only (All types)	**H**
7	G63	10	Hemopoietic only (All types)	**H**
8	G64	15	Hemopoietic only (All types)	**H**
9	G70	44	Hemopoietic only (esp. in NK, CD4, CD8, WB)	**H**
10	G71	32	Hemopoietic only (esp. in T, NK, WB, Thymus)	**H**
11	G72	44	Hemopoietic only (esp. in DC, Mono, Myeloid, WB)	**H**
12	G73	279	Hemopoietic only (All types EXCEPT early Erythroid)	**H**
13	G82	62	Hemopoietic only (esp. in Erythroid, Endoth., FL, BM)	**H**
14	G91	25	Hemopoietic only (esp. in Erythroid, Endoth., FL, BM)	**H**
15	G23	15	Hemopoietic & Neuronal tissues	**H, N**
16	G56	95	Hemopoietic & Neuronal tissues	**H, N**
17	G66	22	Hemopoietic & Neuronal tissues	**H, N**
18	G12	15	Hemopoietic & Neuronal & Smooth muscle	**H, N, SM**
19	G15	12	Hemopoietic & Neuronal & Testis	**H, N, T**
20	G19	24	Hemopoietic (esp. in Mono, Myeloid, WB, FL) & Liver	**H, L**
21	G87	127	Hemopoietic & Testis & Cardiac myocytes	**H, T, CM**
22	G32	45	Hemopoietic (esp. in DC, Mono, Myeloid, WB) & Smooth muscle & Cardiac myocytes	**H, SM, CM**
23	G83	67	Hemopoietic & Trachea & Lung	**H, TR, LNG**
24	G42	58	Hemopoietic (Tonsils) & Tongue & Bronch. Epithel. Cells	**H, TNG, BEC**

Table II-2. Clusters of normal human tissues - Non-hematopoietic (NH) clusters

No.	Cluster	No. of genes	Preferentially expressed in	Symbol
1	G2	52	Pancreas only	**PANC**
2	G9	17	Kidney only	**K**
3	G45	100	Testis only	**T**
4	G10	50	Adrenal gland only	**ADR**
5	G81	37	Placenta only	**PL**
6	G5	19	Placenta & Pituitary, a little in lung, fetal lung	**PL, Pit, LNG**
7	G21	10	Neuronal tissues only	**N**
8	G22	21	Neuronal tissues only	**N**
9	G68	75	Neuronal tissues only	**N**
10	G86	322	Neuronal tissues only	**N**
11	G18	20	Neuronal tissues & Testis	**N, T**
12	G26	15	Neuronal tissues & Testis	**N, T**
13	G27	23	Thyroid & fetal thyroid	**THYR, FHTYR**
14	G89	154	Liver, fetal liver & kidney & fetal lung	
15	G40	46	Lung, fetal lung & Trachea	
16	G39	28	Salivary gland & Trachea & a little in Thalamus	**SAL, TR, THA**
17	G20	30	Low in hematopoietic tissues (highest in adipocytes)	**LIH**
18	G24	36	Low in hematopoietic tissues	**LIH**
19	G25	62	Low in hematopoietic tissues (highest in sm. muscle, heart, thyroid)	**LIH**
20	G35	27	Low in hematopoietic tissues (highest in neuronal)	**LIH**
21	G43	59	Low in hematopoietic tissues	**LIH**
22	G44	15	Low in hematopoietic tissues	**LIH**
23	G46	12	Low in hematopoietic tissues	**LIH**
24	G47	24	Low in hematopoietic tissues (highest in neuronal)	**LIH**
25	G48	43	Low in hematopoietic tissues	**LIH**
26	G49	54	Low in hematopoietic tissues	**LIH**
27	G50	62	Low in hematopoietic tissues	**LIH**
28	G85	196	Low in hematopoietic tissues (highest in muscle)	**LIH**

BM, bone marrow; DC, dendritic cells; FL, fetal liver; LN, lymph node; NK, natural killer cells; WB, whole blood.

Table II-3. List of genes in hematopoietic (H) clusters highly expressed in cancer cells but not in their normal counterparts

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
G12	209935_at	**ATP2C1**	-	-	-	+	-	-	-
	217841_s_at	**PME-1**	+	-	-	-	-	-	-
	204141_at	**TUBB**	+	-	-	-	-	-	-
	gnf1h00312_at	**ANLN**	+	-	-	-	-	-	-
	202219_at	**SLC6A8**	+	-	-	-	-	-	-
	210527_x_at	**TUBA2**	+	-	-	-	-	-	-
	213476_x_at	**TUBB4**	+	-	-	-	-	-	-
	201195_s_at	**SLC7A5**	+	-	-	-	-	-	-
G19	204588_s_at	**SLC7A7**	+	-	-	-	-	-	-
	211429_s_at	**SERPINA1**	+	-	-	-	-	-	-
	202833_s_at	**SERPINA1**	+	-	-	-	-	-	-
	202241_at	**TRIB1**	+	-	-	-	-	-	-
G23	217979_at	**TM4SF13**	+	-	-	-	-	-	-
G28	205174_s_at	**QPCT**	+	-	-	-	-	-	-
G30	205445_at	**PRL**	-	-	-	-	+	-	-
	211003_x_at	**TGM2**	+	-	-	-	-	-	-
	211573_x_at	**TGM2**	+	-	-	-	-	-	-
G32	218051_s_at	**FLJ12442**	-	-	+	-	+	+	+
	209921_at	**SLC7A11**	-	-	+	+	+	+	-
	202619_s_at	**PLOD2**	-	-	+	-	-	-	-
	202381_at	**ADAM9**	+	-	-	-	-	-	-
G42	201015_s_at	**JUP**	+	-	-	-	-	-	-
	208502_s_at	**PITX1**	+	-	-	-	-	-	-
	209260_at	**SFN**	+	-	-	-	-	-	-
	207935_s_at	**KRT13**	+	-	-	-	-	-	-
	204990_s_at	**ITGB4**	+	-	-	-	-	-	-
	202504_at	**TRIM29**	+	-	-	-	-	-	-
	202286_s_at	**TACSTD2**	+	-	-	-	-	-	-
	201820_at	**KRT5**	+	-	-	-	-	-	-

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
	204268_at	**S100A2**	-	-	-	-	-	-	+
G56	204040_at	**RNF144**	+	-	-	-	-	-	-
	213587_s_at	**C7orf32**	+	-	-	-	-	-	-
	212068_s_at	**KIAA0515**	+	-	-	-	-	-	-
	218404_at	**SNX10**	+	-	-	-	-	-	-
	216033_s_at	**FYN**	+	-	-	-	-	-	-
	202206_at	**ARL7**	+	-	-	-	-	-	-
	202208_s_at	**ARL7**	+	-	-	-	-	-	-
	202806_at	**DBN1**	+	-	-	-	-	-	-
	200973_s_at	**TM4SF8**	+	-	-	-	-	-	-
	209185_s_at	**IRS2**	+	-	-	-	-	-	-
G60	205321_at	**EIF2S3**	+	-	-	-	-	-	-
G63	gnf1h04674_at	**C6orf83**	+	-	-	-	-	-	-
	217868_s_at	**DREV1**	+	-	-	-	-	-	-
	213102_at	**ACTR3**	+	-	-	-	-	-	-
	208901_s_at	**TOP1**	+	-	-	-	-	-	-
G64	202848_s_at	**GRK6**	+	-	-	-	-	-	-
	202771_at	**FAM38A**	+	-	-	-	-	-	-
	216237_s_at	**MCM5**	+	-	-	-	-	-	-
	201202_at	**PCNA**	+	-	-	-	-	-	-
G66	gnf1h06906_at	**DKFZp313A2432**	-	-	+	-	-	-	-
	218123_at	**C21orf59**	+	-	-	-	-	-	-
	214870_x_at	**NPIP**	+	-	-	-	-	-	-
	203839_s_at	**ACK1**	+	-	-	-	-	-	-
G70	210140_at	**CST7**	+	-	-	-	-	-	-
	214617_at	**PRF1**	+	-	-	-	-	-	-
G71	207979_s_at	**CD8B1**	+	-	-	-	-	-	-
	211796_s_at	**211796_s_at**	+	-	-	-	-	-	-
G73	210448_s_at	**P2RX5**	+	-	-	-	-	-	-
	206200_s_at	**ANXA11**	+	-	-	-	-	-	-
	205483_s_at	**G1P2**	+	-	-	-	-	-	-
	217984_at	**RNASET2**	+	-	-	-	-	-	-
	209360_s_at	**RUNX1**	+	-	-	-	-	-	-

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
	209282_at	**PRKD2**	+	-	-	-	-	-	-
G73	207540_s_at	**SYK**	+	-	-	-	-	-	-
	204401_at	**KCNN4**	+	-	-	-	-	-	-
	203508_at	**TNFRSF1B**	+	-	-	-	-	-	-
	gnf1h01004_at	**SYTL1**	+	-	-	-	-	-	-
	219202_at	**RHBDL6**	+	-	-	-	-	-	-
	205081_at	**CRIP2**	+	-	-	-	-	-	-
G82	gnf1h03395_at	**gnf1h03395_at**	-	-	-	-	-	-	+
G87	205345_at	**BARD1**	-	-	+	+	-	-	-
	203976_s_at	**CHAF1A**	-	-	+	-	-	-	-
	218542_at	**C10orf3**	-	-	-	+	-	-	-
	211300_s_at	**TP53**	+	-	+	-	-	-	-
	218847_at	**IMP-2**	+	-	-	-	-	-	-
	204249_s_at	**LMO2**	+	-	-	-	-	-	-
	39729_at	**PRDX2**	+	-	-	-	-	-	-
	gnf1h01357_s_at	**APOBEC3F**	+	-	-	-	-	-	-
	gnf1h00835_at	**PHACS**	+	-	-	-	-	-	-
	211543_s_at	**GRK6**	+	-	-	-	-	-	-
	209208_at	**MPDU1**	+	-	-	-	-	-	-
	218726_at	**DKFZp762E1312**	+	-	-	-	-	-	-
	222037_at	**AI859865**	+	-	-	-	-	-	-
	222036_s_at	**AI859865**	+	-	-	-	-	-	-
	201291_s_at	**TOP2A**	+	-	-	-	-	-	-
	201890_at	**RRM2**	+	-	-	-	-	-	-
	204244_s_at	**ASK**	+	-	-	-	-	-	-
	202580_x_at	**FOXM1**	+	-	-	-	-	-	-
	212330_at	**TFDP1**	+	-	-	-	-	-	-
	202338_at	**TK1**	+	-	-	-	-	-	-
	202095_s_at	**BIRC5**	+	-	-	-	-	-	-
	202107_s_at	**MCM2**	+	-	-	-	-	-	-
	201664_at	**SMC4L1**	+	-	-	-	-	-	-
	201663_s_at	**SMC4L1**	+	-	-	-	-	-	-
	201014_s_at	**PAICS**	+	-	-	-	-	-	-

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
	201088_at	**KPNA2**	+	-	-	-	-	-	-
	204170_s_at	**CKS2**	+	-	-	-	-	-	-
G87	204825_at	**MELK**	+	-	-	-	-	-	-
	205436_s_at	**H2AFX**	+	-	-	-	-	-	-
	218009_s_at	**PRC1**	+	-	-	-	-	-	-
	gnf1h00157_at	**HSPC150**	+	-	-	-	-	-	-
	gnf1h05869_s_at	**gnf1h05869_s_at**	+	-	-	-	-	-	-
	209153_s_at	**TCF3**	+	-	-	-	-	-	-
	221932_s_at	**C14orf87**	+	-	-	-	-	-	-
	210983_s_at	**MCM7**	+	-	-	-	-	-	-
	208795_s_at	**MCM7**	+	-	-	-	-	-	-
	208691_at	**TFRC**	+	-	-	-	-	-	-
	207332_s_at	**TFRC**	+	-	-	-	-	-	-
	202870_s_at	**CDC20**	+	-	-	-	-	-	-
	202779_s_at	**UBE2S**	+	-	-	-	-	-	-
	207165_at	**HMMR**	+	-	-	-	-	-	-
	201292_at	**TOP2A**	+	-	-	-	-	-	-
	202503_s_at	**KIAA0101**	+	-	-	-	-	-	-
	202589_at	**TYMS**	+	-	-	-	-	-	-
	204767_s_at	**FEN1**	+	-	-	-	-	-	-
	gnf1h00130_at	**UHRF1**	+	-	-	-	-	-	-
	gnf1h00245_s_at	**MRPL37**	+	-	-	-	-	-	-

Table II-4. List of genes in non-hematopoietic (NH) clusters highly expressed in cancer cells but not in their normal counterparts

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
Pancreas									
G2	216470_x_at	**216470_x_at**	-	-	-	-	-	-	+
Kidney									
G9	203559_s_at	**ABP1**	+	-	+	-	-	-	-
	207434_s_at	**FXYD2**	-	+	-	-	-	-	-
	205674_x_at	**FXYD2**	-	+	-	-	-	-	-
	gnf1h02291_at	**gnf1h02291_at**	-	+	-	-	-	-	-
Adrenal									
G10	205633_s_at	**ALAS1**	+	-	-	-	-	-	-
	203647_s_at	**FDX1**	+	-	-	-	-	-	-
	208928_at	**POR**	+	-	-	-	-	-	-
	208161_s_at	**ABCC3**	+	-	-	-	-	-	-
	216609_at	**TXN**	-	+	-	-	+	+	-
	207813_s_at	**FDXR**	-	-	+	-	-	-	-
	gnf1h06269_at	**ATP6V1C2**	-	-	-	-	-	+	-
	209560_s_at	**DLK1**	-	-	-	-	-	-	+
Thyroid									
G27	205350_at	**CRABP1**	-	+	-	-	-	-	-
	204259_at	**MMP7**	-	-	+	-	-	-	-
Neuronal									
G21	206140_at	**LHX2**	-	-	+	-	-	-	-
G68	205691_at	**SYNGR3**	-	-	-	+	-	+	+
	201662_s_at	**ACSL3**	-	+	+	-	-	-	+
	gnf1h00805_at	**SCOC**	-	+	+	-	-	-	-
	219170_at	**FSD1**	-	-	-	+	-	-	-
	205210_s_at	**FGF13**	-	-	-	-	-	-	+
	201462_at	**SCRN1**	+	-	-	-	-	-	-
	212221_x_at	**IDS**	+	-	-	-	-	-	-
	212223_at	**IDS**	+	-	-	-	-	-	-
G68	213135_at	**TIAM1**	+	-	-	-	-	-	-
	219549_s_at	**RTN3**	+	-	-	-	-	-	-
	200623_s_at	**CALM3**	+	-	-	-	-	-	-
G86	202260_s_at	**STXBP1**	+	-	-	+	-	+	-
	204073_s_at	**C11orf9**	+	-	+	-	-	-	+
	201387_s_at	**UCHL1**	-	-	+	+	-	+	-
	204730_at	**RIMS3**	-	-	-	+	+	+	-
	219236_at	**PAQR6**	-	-	-	+	+	-	-
	216963_s_at	**GAP43**	-	-	-	+	-	+	-
	215116_s_at	**DNM1**	+	-	-	-	-	-	+
	214023_x_at	**MGC8685**	+	-	-	-	-	-	+

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
	205970_at	**MT3**	-	-	-	+	-	-	-
	209598_at	**PNMA2**	-	+	-	-	-	-	-
	203069_at	**SV2A**	-	+	-	-	-	-	-
	gnf1h08659_at	**UNC13C**	-	-	+	-	-	-	-
	209470_s_at	**GPM6A**	-	-	+	-	-	-	-
	209469_at	**GPM6A**	-	-	+	-	-	-	-
	205399_at	**DCAMKL1**	+	-	-	-	-	-	-
	202517_at	**CRMP1**	+	-	-	-	-	-	-
	204540_at	**EEF1A2**	+	-	-	-	-	-	-
	204584_at	**L1CAM**	+	-	-	-	-	-	-
	218417_s_at	**FLJ20489**	+	-	-	-	-	-	-
	212559_at	**PRKAR1B**	+	-	-	-	-	-	-
	204724_s_at	**COL9A3**	+	-	-	-	-	-	-
	218952_at	**PCSK1N**	+	-	-	-	-	-	-
	gnf1h07687_at	**EPHA4**	+	-	-	-	-	-	-
	203961_at	**NEBL**	+	-	-	-	-	-	-
	203955_at	**KIAA0649**	+	-	-	-	-	-	-
	212233_at	**MAP1B**	-	-	-	-	-	-	+
	213338_at	**RIS1**	-	-	-	-	-	-	+
	206453_s_at	**NDRG2**	-	-	-	-	-	-	+
	219196_at	**SCG3**	-	-	-	-	-	-	+
	205625_s_at	**CALB1**	-	-	-	-	-	-	+
Testis									
G45	220110_s_at	**NXF3**	-	-	+	-	-	-	-
	214296_x_at	**IMAGE:4215339**	-	-	-	+	-	-	-
	207739_s_at	**GAGE5**	-	-	-	-	-	-	+
	206640_x_at	**GAGE5**	-	-	-	-	-	-	+
Placenta									
G81	219424_at	**EBI3**	-	-	+	+	-	-	-
	208257_x_at	**PSG1**	+	-	-	-	-	-	-
	204830_x_at	**PSG4**	+	-	-	-	-	-	-
	205602_x_at	**PSG7**	+	-	-	-	-	-	-
Neuronal and testis									
G18	213479_at	**NPTX2**	+	-	-	-	-	-	-
	209343_at	**EFHD1**	-	-	-	-	-	-	+
	gnf1h05957_at	**gnf1h05957_at**	-	-	-	-	-	-	+
Salivary gland and trachea									
G39	206224_at	**CST1**	+	-	-	-	-	-	-
	208555_x_at	**CST2**	+	-	-	-	-	-	-
Lung and trachea									
G40	209270_at	**LAMB3**	+	-	-	-	-	-	-
	214651_s_at	**HOXA9**	+	-	-	-	-	-	-
	gnf1h01731_s_at	**gnf1h01731_s_at**	+	-	-	-	-	-	-
	203108_at	**RAI3**	+	-	-	-	-	-	-

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
	205366_s_at	**HOXB6**	+	-	-	-	-	-	-
	205749_at	**CYP1A1**	-	-	-	-	-	-	+
Placenta, pituitary and lung									
G5	207770_x_at	**CSH2**	-	-	-	+	-	-	-
Liver, kidney and fetal lung									
G89	201674_s_at	**AKAP1**	+	+	+	-	+	+	-
	gnf1h01169_at	**WDR34**	+	+	+	-	-	-	-
	202740_at	**ACY1**	+	-	-	+	-	-	+
	204044_at	**QPRT**	-	-	+	+	-	-	+
	216381_x_at	**AKR7A3**	+	-	-	-	-	+	-
G89	209081_s_at	**COL18A1**	+	-	-	-	-	-	-
	205774_at	**F12**	+	-	-	-	-	-	-
	217188_s_at	**C14orf1**	-	-	+	-	-	-	-
	205208_at	**FTHFD**	-	-	+	-	-	-	-
	206754_s_at	**CYP2B6**	-	-	+	-	-	-	-
	209975_at	**CYP2E1**	-	-	-	+	-	-	-
	205650_s_at	**FGA**	-	-	-	-	+	-	-
	219733_s_at	**SLC27A5**	-	-	-	-	-	+	-
	205943_at	**TDO2**	-	-	-	-	-	-	+
Low in hematopoietic (LIH)									
G20	200832_s_at	**SCD**	+	+	+	+	+	+	+
	209146_at	**SC4MOL**	+	+	+	+	+	-	+
	gnf1h01546_s_at	**Hs.523212**	+	+	-	-	-	+	+
	212186_at	**ACACA**	-	+	+	-	-	+	+
	208963_x_at	**FADS1**	-	+	+	-	-	-	+
	202540_s_at	**HMGCR**	-	+	+	-	-	-	+
	211162_x_at	**SCD**	-	-	+	-	-	-	-
	211708_s_at	**SCD**	-	-	+	-	-	-	-
	205498_at	**GHR**	-	-	+	-	-	-	-
	200831_s_at	**SCD**	+	-	-	-	-	-	-
	210830_s_at	**PON2**	+	-	-	-	-	-	-
	217776_at	**RDH11**	+	-	-	-	-	-	-
	200947_s_at	**GLUD1**	+	-	-	-	-	-	-
G24	205742_at	**TNNI3**	-	-	-	-	-	-	+
G25	219188_s_at	**LRP16**	-	-	-	+	+	-	-
	205177_at	**TNNI1**	-	-	-	+	-	-	-
	206353_at	**COX6A2**	-	-	-	-	+	-	-
	213201_s_at	**TNNT1**	-	-	-	-	-	-	+
G35	207169_x_at	**DDR1**	+	-	-	-	-	-	-
	215807_s_at	**PLXNB1**	+	-	-	-	-	-	-
	219305_x_at	**FBXO2**	+	-	-	-	-	-	-
G35	213050_at	**COBL**	+	-	-	-	-	-	-
	204447_at	**ProSAPiP1**	-	-	-	+	-	-	-

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
	214797_s_at	**PCTK3**	-	-	-	+	-	-	-
G43	221577_x_at	**GDF15**	+	-	-	-	-	-	+
	209008_x_at	**KRT8**	+	-	-	-	-	-	-
	201596_x_at	**KRT18**	+	-	-	-	-	-	-
	200636_s_at	**PTPRF**	+	-	-	-	-	-	-
	210715_s_at	**SPINT2**	+	-	-	-	-	-	-
	217744_s_at	**PERP**	+	-	-	-	-	-	-
	36711_at	**36711_at**	+	-	-	-	-	-	-
	gnf1h01008_at	**JUB**	+	-	-	-	-	-	-
	gnf1h06417_s_at	**ZD52F10**	+	-	-	-	-	-	-
	gnf1h11118_x_at	**gnf1h11118_x_at**	+	-	-	-	-	-	-
	218963_s_at	**KRT23**	+	-	-	-	-	-	-
	212444_at	**RAI3**	+	-	-	-	-	-	-
	208190_s_at	**LISCH7**	+	-	-	-	-	-	-
	203453_at	**SCNN1A**	+	-	-	-	-	-	-
	203407_at	**PPL**	+	-	-	-	-	-	-
	202826_at	**SPINT1**	+	-	-	-	-	-	-
	202790_at	**GABARAP**	+	-	-	-	-	-	-
	201650_at	**KRT19**	+	-	-	-	-	-	-
	201474_s_at	**ITGA3**	+	-	-	-	-	-	-
	201428_at	**CLDN4**	+	-	-	-	-	-	-
	203954_x_at	**CLDN3**	+	-	-	-	-	-	-
	200606_at	**DSP**	+	-	-	-	-	-	-
G44	204345_at	**COL16A1**	-	-	+	-	-	-	-
	216620_s_at	**ARHGEF10**	-	+	-	-	-	-	-
	219926_at	**POPDC3**	-	-	-	-	-	-	+
	201334_s_at	**ARHGEF12**	+	-	-	-	-	-	-
G47	212094_at	**PEG10**	-	-	+	+	-	+	-
	203029_s_at	**PTPRN2**	+	+	-	-	-	-	-
	210794_s_at	**MEG3**	-	-	-	-	-	-	+
	212062_at	**ATP9A**	+	-	-	-	-	-	-
G48	202238_s_at	**NNMT**	-	-	-	-	-	-	+
	202718_at	**IGFBP2**	+	-	-	-	-	-	-
	203423_at	**RBP1**	+	-	-	-	-	-	-
G49	212110_at	**SLC39A14**	+	-	-	+	+	+	+
	201564_s_at	**FSCN1**	-	-	+	-	-	-	+
	203786_s_at	**TPD52L1**	-	-	+	-	-	-	+
	221538_s_at	**DKFZp564A176**	+	-	-	-	-	-	+
	202067_s_at	**LDLR**	-	-	-	-	-	-	+
	gnf1h00310_at	**ACAS2**	-	-	-	-	-	-	+
	201889_at	**FAM3C**	+	-	-	-	-	-	-
	208029_s_at	**LAPTM4B**	+	-	-	-	-	-	-
	202976_s_at	**RHOBTB3**	+	-	-	-	-	-	-
	217849_s_at	**CDC42BPB**	+	-	-	-	-	-	-
	202371_at	**FLJ21174**	+	-	-	-	-	-	-

Cluster	Probe set	Name	SW480	Molt4	721	Raji	Daudi	HL-60	K562
G50	216215_s_at	**RBM9**	-	-	-	+	+	+	+
	213423_x_at	**TUSC3**	-	-	-	-	-	-	+
	200884_at	**CKB**	-	-	-	-	-	-	+
	209094_at	**DDAH1**	+	-	-	-	-	-	-
	213069_at	**HEG**	+	-	-	-	-	-	-
	203962_s_at	**NEBL**	+	-	-	-	-	-	-
	202936_s_at	**SOX9**	+	-	-	-	-	-	-
	202935_s_at	**SOX9**	+	-	-	-	-	-	-
	202458_at	**SPUVE**	+	-	-	-	-	-	-
G50	200602_at	**APP**	+	-	-	-	-	-	-
G85	209344_at	**TPM4**	+	-	+	-	-	-	-
	202733_at	**P4HA2**	+	-	-	-	-	-	+
	201125_s_at	**ITGB5**	+	-	-	-	-	-	+
	214020_x_at	**ITGB5**	+	-	-	-	-	-	+
	209262_s_at	**NR2F6**	-	-	-	+	-	+	-
	209121_x_at	**NR2F2**	-	-	+	-	-	-	+
	202620_s_at	**PLOD2**	-	-	+	-	-	-	-
	222288_at	**Hs.130526**	-	+	-	-	-	-	-
	204518_s_at	**PPIC**	-	-	-	-	-	-	+
	213800_at	**HF1**	-	-	-	-	-	-	+
	200755_s_at	**CALU**	-	-	-	-	-	-	+
	208712_at	**CCND1**	+	-	-	-	-	-	-
	201983_s_at	**EGFR**	+	-	-	-	-	-	-
	219922_s_at	**LTBP3**	+	-	-	-	-	-	-
	212698_s_at	**SEPT10**	+	-	-	-	-	-	-
	202627_s_at	**SERPINE1**	+	-	-	-	-	-	-
	202628_s_at	**SERPINE1**	+	-	-	-	-	-	-
	203438_at	**STC2**	+	-	-	-	-	-	-
	204306_s_at	**CD151**	+	-	-	-	-	-	-
	203180_at	**ALDH1A3**	+	-	-	-	-	-	-
	202949_s_at	**FHL2**	+	-	-	-	-	-	-

Table II-5. List of genes in hematopoietic (H) and nonhematopoietic (NH) clusters highly expressed in human leukemias but not in any normal H tissue

			T-ALL	B-ALL						AML					
Cluster	Probe set	Name	MLL*		MLL*	BCR/ABL	E2A/PBX1	TEL/AML1	HD50	Others	MLL*	PML/RARA	AML1/ETO	CBFβ/MYH11	M7	Others
H																
G30	205445_at	PRL	-	-	-	-	-	-	-	-	+	-	-	-	-	-
G87	205345_at	BARD1	-	+	-	-	+	-	-	-	-	-	-	-	-	-
NH																
Kidney																
G9	203913_s_at	HPGD	+	-	-	-	-	-	-	-	-	-	-	-	-	-
Adrenal																
G10	209560_s_at	DLK1	-	-	-	-	-	-	-	-	+	-	-	-	-	-
	205083_at	AOX1	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Tetsis																
G45	220230_s_at	CYB5R2	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Placenta																
G81	221710_x_at	FLJ10647	-	-	-	-	-	-	-	-	-	+	-	-	-	-
	213094_at	GPR126	-	-	-	-	-	-	-	-	+	-	-	-	-	-
Neuronal																
G68	208820_at	PTK2	-	-	-	-	-	+	-	-	-	-	-	-	-	-
	205352_at	SERPINI1	-	-	-	-	-	-	-	-	-	-	-	-	+	-
G86	202517_at	CRMP1	-	-	-	-	+	+	-	-	-	-	-	-	-	-
	203961_at	NEBL	+	-	-	-	-	-	-	-	-	-	-	-	-	-
	212358_at	CLIPR-59	+	-	-	-	-	-	-	-	-	-	-	-	-	-
	209168_at	GPM6B	-	-	+	-	-	-	-	-	-	-	-	-	-	-
	220448_at	KCNK12	-	-	+	-	-	-	-	-	-	-	-	-	-	-
	203999_at	SYT1	-	-	-	-	+	-	-	-	-	-	-	-	-	-
	218613_at	EFA6R	-	-	-	-	-	+	-	-	-	-	-	-	-	-
	215116_s_at	DNM1	-	-	-	-	-	-	-	-	-	+	-	-	-	-
	206135_at	ST18	-	-	-	-	-	-	-	-	-	-	-	+	-	-
Neuronal and testis																
G18	213479_at	NPTX2	-	-	+	-	-	-	-	-	-	-	-	-	-	-

	213791_at	PENK	-	-	+	-	-	-	-	-	-	-	-	-	-	-
	209781_s_at	KHDRBS3	-	-	-	-	-	+	-	-	+	-	-	-	-	-
G26	204165_at	WASF1	-	-	+	-	+	-	+	-	-	-	-	-	-	-
	205109_s_at	ARHGEF4	-	-	-	-	-	+	-	-	-	-	-	-	-	-
Liver, kidney and fetal lung																
G89	204561_x_at	APOC2	-	-	-	-	-	-	-	-	-	-	-	-	+	-
Low in hematopoietic (LIH)																
G20	200832_s_at	SCD	+	+	-	-	-	-	-	-	-	-	-	-	-	-
	202540_s_at	HMGCR	+	-	-	-	-	-	-	-	-	-	-	-	-	-
	209146_at	SC4MOL	+	-	-	-	-	-	-	-	-	-	-	-	-	-
G24	210298_x_at	FHL1	-	-	-	+	-	+	+	-	-	-	-	-	+	-
G44	212148_at	PBX1	-	-	-	-	+	-	-	-	-	-	-	-	+	-
G47	212062_at	ATP9A	+	-	-	-	-	-	-	-	-	-	-	-	-	-
	210794_s_at	MEG3	-	-	-	-	-	-	-	-	-	+	-	-	-	-
G48	212154_at	SDC2	-	-	+	-	-	+	-	-	-	-	-	-	-	-
	212158_at	SDC2	-	-	-	-	-	+	-	-	-	-	-	-	-	-
G49	201564_s_at	FSCN1	-	-	-	-	-	-	-	-	-	+	-	-	-	-
G50	207030_s_at	CSRP2	-	-	+	+	-	+	-	-	-	-	-	-	-	-
	203131_at	PDGFRA	-	-	-	-	-	-	-	+	-	-	-	-	-	-
	203962_s_at	NEBL	+	-	-	-	-	-	-	-	-	-	-	-	-	-
G85	209344_at	TPM4	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	212013_at	D2S448	-	+	-	+	+	+	+	+	-	+	-	-	-	-
	212488_at	Hs.433695	-	-	-	+	+	+	+	+	-	-	-	-	-	-
	222288_at	Hs.130526	+	-	-	-	-	-	-	-	-	-	-	-	-	-
	218468_s_at	GREM1	-	-	+	-	-	-	-	-	-	-	-	-	-	-
	202733_at	P4HA2	-	-	-	-	+	-	-	-	-	-	-	-	-	-

Table II-6. Genes in H clusters that are overexpressed in SW480 and have a role in human cancer

	Gene Function						
Up-regulated in human cancer	Transcription factor or oncogene	Cytoskeleton	Adhesion, invasiveness, angiogenesis	Proliferation	DNA repair, replication	Transport	Antiapoptosis
QPCT	FYN	TUBA2	ADAM9	SFN	TOP1	SLC7A5	SERPINA1
SERPINA1	TP53	TUBB	ITGB4	IRS2	FEN1	SLC7A7	TYMS
ADAM9	LMO2	TUBB4	TM4SF8	MCM2	TOP2A		BIRC5
JUP	TCF3	ACTR3	HMMR	MCM4	RRM2		
TACSTD2	RUNX1			MCM5	ASK		
S100A2				MCM7	TK1		
RUNX1				PCNA	H2AFX		
FEN1				TFDP1			
TYMS				SMC4L1			
HMMR				CKS2			
TUBB				PRC1			
SLC7A5				CDC20			
TCF3				UHRF1			
BIRC5							
ITGB4							
CDC20							
RRM2							

Table II-7. List of genes in NH clusters that are overexpressed in cancer cell lines and have a role in human cancer

Up-regulated in human cancer		Gene function						
		Transcription factor or oncogene	Cytoskeleton	Adhesion, invasiveness,angiogenesis	Proliferation	Transport	Lipid or drug metabolism	Anti-apoptosis
SCD	IGFBP2	LHX2	DSP	MMP7	LRP16	SCNN1A	SCD	PLXNB1
ACACA	RAI3	HOXA9	TPM4	PLXNB1	CCND1		ACACA	PEG10
CRABP1	FSCN1	HOXB6		EPHA4	PLXNB1		NNMT	APP
DDR1	LAPTM4B	LAPTM4B		GDF15	RAI3			EGFR
PLXNB1	SOX9	SOX9		LAMB3	PEG10			SERPINE1
GDF15	SPUVE	FHL2		ITGA3	LAPTM4B			MT3
HOXA9	APP	EEF1A2		ITGB5	EGFR			
HOXB6	P4HA2			IGFBP2				
KRT8	TIAM1			FSCN1				
KRT18	TPM4			TIAM1				
KRT19	SERPINE1			SERPINE1				
GAGE5	STC2			CD151				
ITGB5	CD151			L1CAM				
SCNN1A	FHL2			EPHA4				
PTPRF	MT3			COL18A1				
DSP	EEF1A2							
PEG10	L1CAM							
NNMT	EPHA4							
	COL18A1							