

Microarray Data Analysis and Pathway Visualization of Murine Cochlear Ear Hair Cell Differentiation

Eric S. Allen[‡] & Kris Barbee[‡]

[‡]From the University of California San Diego, Department of Bioengineering

Submitted on December 4th, 2003

Analysis of a gene expression study related to cochlear ear hair cell differentiation experiments has illuminated many challenges with microarray data analysis and pathway modeling. The methods used to analyze microarray data by many groups may lead to inaccurate interpretations and misleading conclusions. In addition to this microarray analysis, we attempted to visualize a pathway showing the roles and interactions of important gene products in this differentiation process. This was performed with the aid of a software package, PathwayAssist, which allowed for the integration of disparate data sources and produced a graphical output that described portions of the pathways involved in the aforementioned differentiation process.

INTRODUCTION

The systems biology approach is showing its usefulness in an ever-expanding number of areas. While beginning as a rather academic pursuit, it is now being used to research the mechanisms of human disease and both dangerous and economically useful microorganisms. This paper will examine the use of the systems biology approach in the areas of tissue engineering and stem cell research. First, the larger subject of tissue engineering will be discussed; along with the role that systems biology can play in this field. Then, a specific microarray study that deals with the differentiation of cochlear ear hair cells will be analyzed. The ability to control this process could provide a tissue engineering based remedy for hearing loss. Next, a pathway visualization tool will be demonstrated that has the ability to build and visualize pathways based on a wide variety of biological data. Tools such as this have the potential to help researchers understand incredibly complex biological systems, and are excellent examples of the power of the systems biology approach.

Systems Biology in a Tissue Engineering and Cell-based Therapy Context

The term tissue engineering, a relatively new interdisciplinary field, describes the merger of the principles and methods of engineering with those known in the life sciences that makes possible the development of substitute tissue in order to replace that which is damaged, diseased or incompetent [1]. The primary purpose of tissue engineering is to produce natural tissues and organs to replace those that are damaged or failing. This is to be done using living cells, preferably from the individual in need of the transplant, in order to produce tissues and organs that will not be recognized as foreign material to the patient's immune system [2]. This would help ensure acceptance of the transplant, thus eliminating the need for potentially harmful immunosuppressant drugs. Furthermore, a system that could exploit the body's ability to heal

itself could drastically alter the economic state of health care by significantly reducing the cost of expensive procedures such as major organ transplants.

Cell based therapy is another emerging medical technology that involves the transplantation of cells into the patient in order to restore a lost function, repair damaged or diseased tissue, or regenerate a failing organ. One major difference between this approach and that of tissue engineering is that cell-based therapy typically involves an injection of specialized or modified cells directly into the patient at the diseased or damaged site whereas most tissue engineering methods currently employ in-vitro methods to produce or begin producing the necessary tissue prior to implantation. This difference is advantageous in that direct cell implantation techniques are much simpler, less invasive, and often more versatile than approaches requiring an in-vitro developmental stage and subsequent transplantation of an engineered tissue.

Since organs and tissues are comprised of cells, it is intuitive that the type of cells used is of extreme importance in tissue engineering and cell-based therapy. Currently, one of the most popular sources of cells is the embryonic stem cell due to its versatility, capacity for self-renewal, and ability to generate a multitude of other cell types through differentiation. Directing this differentiation requires that the appropriate biological, physical, and chemical stimuli be present. Furthermore, since the examination of cell morphology and the detection of cell specific markers have proven to be inadequate in determining whether the appropriate phenotype of the desired cell type has been achieved, the gene expression profiles of the differentiated cells must be analyzed to ensure that they match those of the native cells [3].

Based on experimental evidence, it appears that the aforementioned cues necessary for the appropriate differentiation of a stem cell into a particular lineage will not be completely recognized without first developing an understanding of the underlying developmental pathways at the genetic level. This will most certainly require the identification of a multitude of gene networks and thus demands that a systems biology approach be taken. One way to accomplish this task is to use currently available information about protein functions and interactions as well as cellular pathways that are scattered throughout various databases and publications to construct such networks. This approach will focus not just on the functions of gene products but also on the relationships between them and the roles of their coordinated activities.

In this study, a commercially available software package, PathwayAssist, was used to perform these complex tasks on a specific differentiation process. The goal was to combine various sets of data with that in the program's database in order to discover key relationships and interactions involved in the differentiation of embryonic stem cells into cochlear hair cells. One published study concerning this process used expression levels derived from gene microarrays to predict the mechanisms responsible for this conversion [4]. Another study experimented with various growth factors in an attempt to produce a desired phenotype [5]. The motivation to conduct the following investigation stems from the fact that neither of the aforementioned studies attempted to analyze the underlying pathways from multiple perspectives.

MATERIALS AND METHODS

The gene expression study [4] and corresponding data were chosen from the collection at the NCBI GEO (Gene Expression Omnibus) site. Two data sets (GDS45 and GDS51) were associated with the reference series GSE36, one for each of the two Affymetrix mouse chips (MU11K A and B). Only the first of the two time courses was analyzed, as it was more complete (11 days vs. 4 days in time series). Raw absolute value data was globally scaled to a mean of 500. Then each non-negative value was converted to log base 2. Values below zero were given a log value of "NA". Next, log fold changes were calculated. In cases where either of the compared values had a "NA" log value, then the log fold change was also made "NA". The data sets did not have replicates, so significance of change tests such as ANOVA or t-tests were not performed.

Next, the web-based program SOURCE [6] from the Stanford Microarray Database group was used to find Unigene and Genbank names associated with each Affymetrix feature, based on the provided Genbank accession. This was useful later when analyzing the clustering results. Then, the data was imported into EPCLUST for hierarchical and K-means clustering. We used the Spearman coefficient for distance estimation for hierarchical clustering, and used K=10 for K-means clustering. Note that the clustering results are not shown or discussed further in this paper, because most of the genes discussed in the Rivolta et al study either had expression values of "NA" due to negative Affymetrix absolute value measurements, or the genes had log₂ fold changes below our cutoffs (between 1 and -1).

For the pathway building and visualizations, we used a demo copy of the program PathwayAssist, distributed by Stratagene. Of the programs we evaluated for pathway creation and manipulation, PathwayAssist was the easiest to use and provided the most useful features. PathwayAssist uses LocusLink IDs (retrieved by SOURCE) to relate genes to data in its own literature-derived database. In addition, the software allowed simultaneous visualization of the expression data with other types of data on a pathway diagram. The features of this product are discussed in more detail in the Results section.

RESULTS

Analysis of Microarray Data

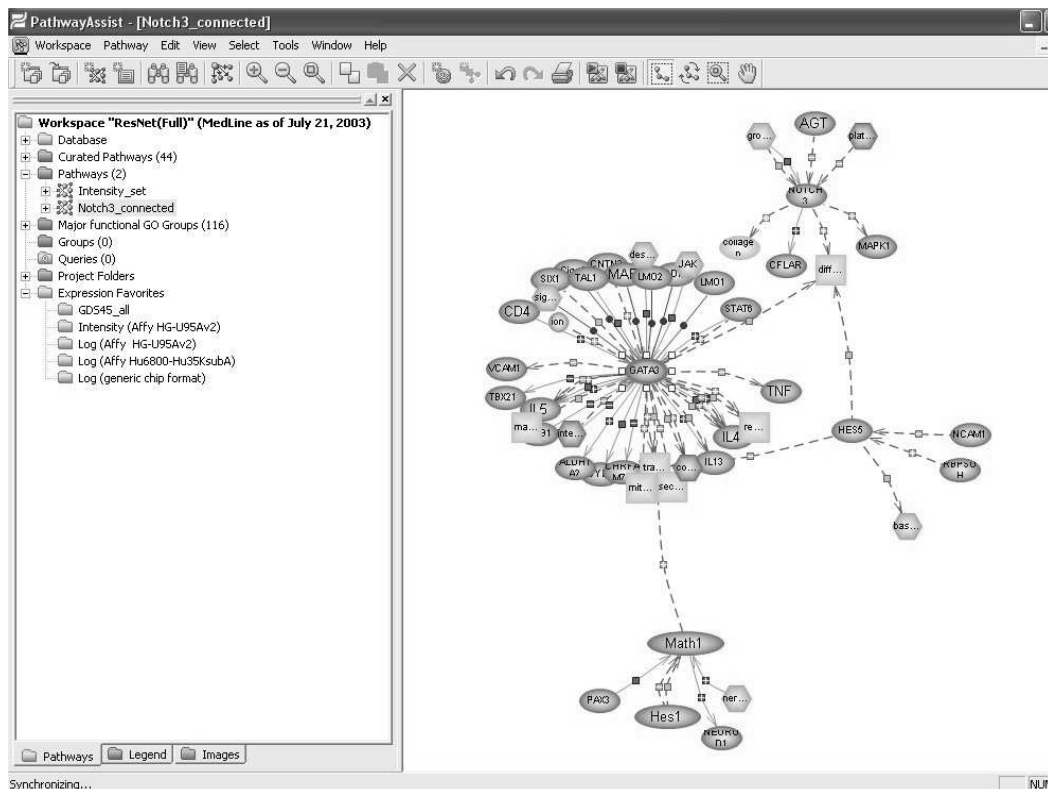
We found several issues with the microarray data results as reported by Rivolta et al. First, replicate readings were not performed, so the expression level measurements were less reliable, and significance tests such as t-tests and ANOVA could not be performed. Although a second time course was performed, it was using a different RNA sample, and thus really wasn't measuring the same thing. Second, when the average difference was <20 for any particular gene, Rivolta et al arbitrarily set all of these values to 20. This had the effect of skewing the fold change results, and made many fold change comparisons look quite large and thus significant. Our team chose to disregard genes that had negative average differences, since we couldn't assign them a value with any confidence level. Third, some of the gene names used by Rivolta et al were apparently not the most commonly used gene names, and were hard to identify. This is actually a very common problem, not just with this particular study.

Table 1 in the Appendix lists several of the genes that were declared in the Rivolta study to have had changes in gene expression levels. We have compared our team's expression assessment to Rivolta et al's, and found many disagreements.

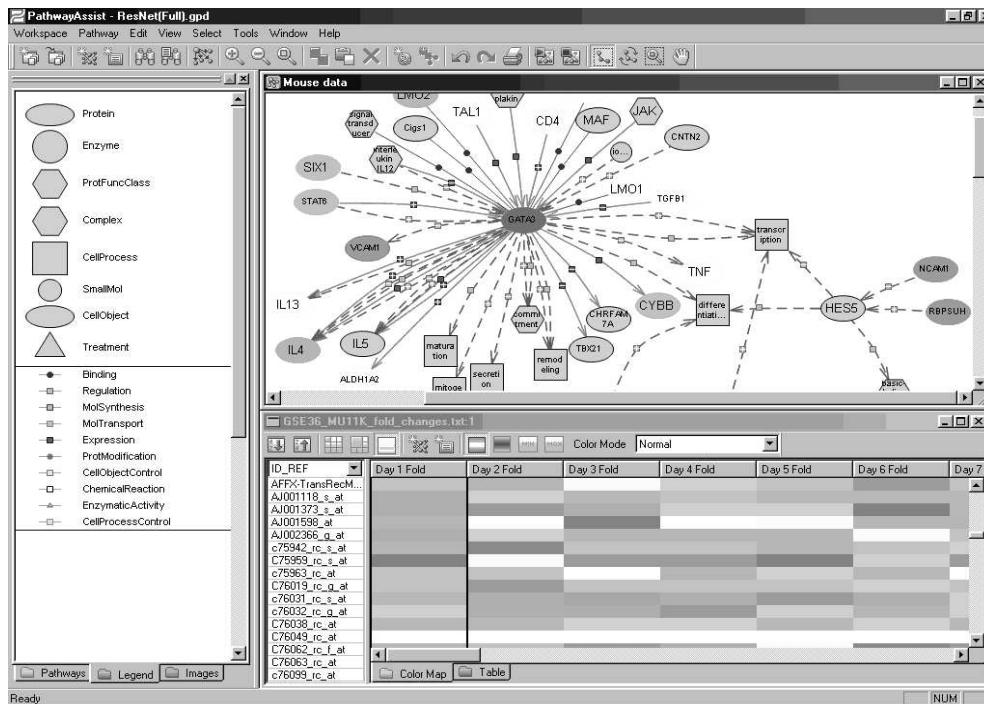
Building and Visualizing Pathways with PathwayAssist

PathwayAssist is a Windows-based application that allows researchers to visually build pathways using a variety of methods. The software comes with a built-in resource named ResNet, which is a database of molecular interactions based on natural language processing of scientific abstracts in PubMed [7]. Using ResNet, a researcher can simply drag his favorite gene product onto a new pathway diagram, and build a pathway using well-known interactions discussed in existing literature. Additionally, the software can easily import data from Kegg, DIP, and BIND. Other formats can be imported if the data is put into a special tab-delimited format. Expression data is also easily imported, with a minimal amount of data preprocessing. Imported expression data can be visually displayed on an existing pathway diagram by showing different shades of green/red depending on the fold change of expression. Several screen shots of PathwayAssist are included in the following pages, as well as a discussion of how we used the product.

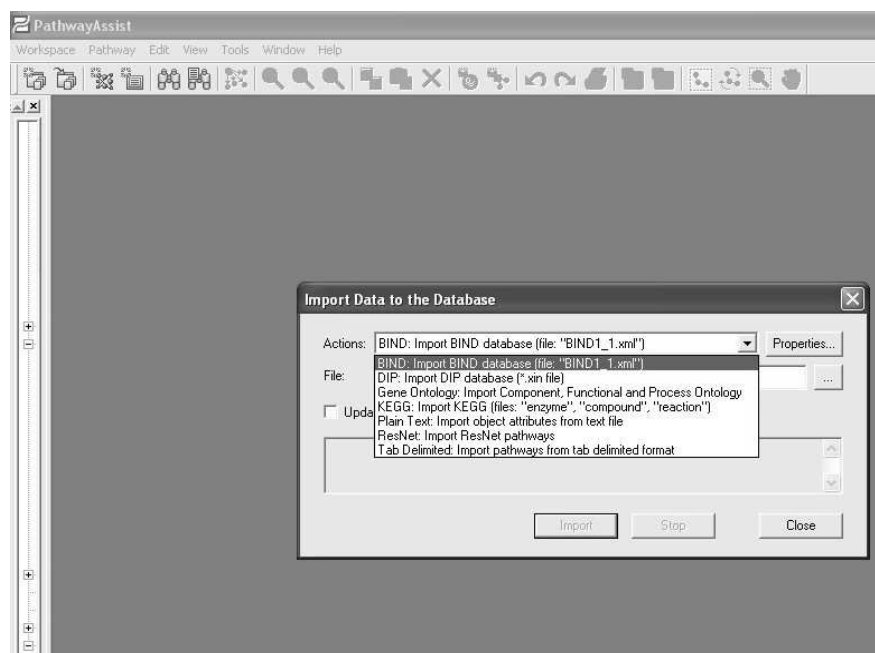
In the screen shot below, a sample pathway from the Rivolta paper has been assembled. On the left is a hierarchical tree, where different types of biological data and associated diagrams can be accessed. On the right is a pathway diagram. The most important objects are proteins shown in red circles and interactions shown as dotted lines.



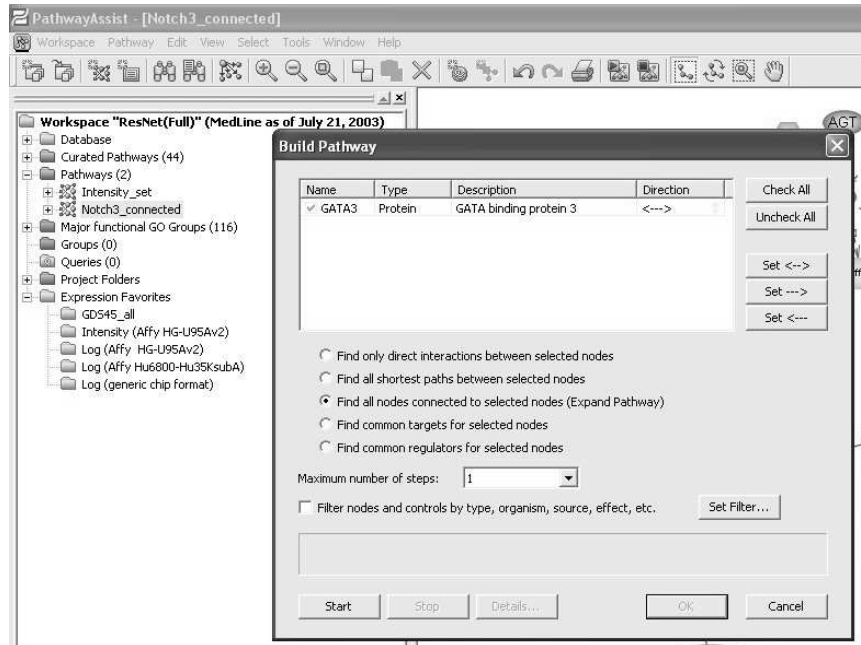
The pathway diagram below has incorporated expression data from the Rivolta et al study. Genes that are up-regulated are shown in shades of green, while genes that are down-regulated are shown in shades of red. When a match can be found for a gene in the expression data and a gene in the pathway diagram, the expression level is indicated. Otherwise, the color gray is used. Also, on the left hand side is a legend that explains what each of the object shapes mean.



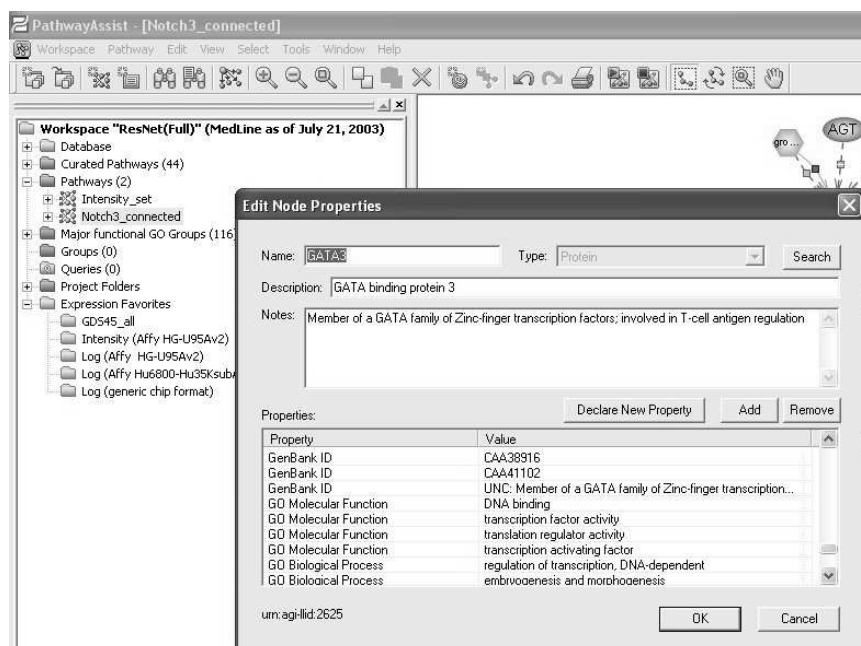
This dialog box below shows how external databases are imported. The example below has BIND highlighted, but some other options are DIP, GO, KEGG, and tab-delimited.



This dialog box shows several methods of automatically building pathways. The option selected “Find all nodes connected to selected nodes (Expand Pathway)”, will search the current pathway database and ResNet for interactions with the currently selected node, and add them to the pathway. Other pathway building options are also available in the dialog box.



The dialog box below shows the “Properties” window, which is available for all nodes and links shown in the pathway diagram, and for all other biological objects in the database. Note that a single object can have multiple external database IDs, GO terms, and synonyms, as well as other properties.



Using Pathway Assist to Study Cochlear Ear Hair Cell Differentiation

We tested the functionality of PathwayAssist by creating a new pathway containing several of the proteins mentioned in the Rivolta and Li et al studies. The result was a graphical summary of some of the proteins involved, those proteins and complexes with which they interact, and the cellular processes for which they are, at least in part, responsible. This offered us a fair amount of insight into the complexity of this particular differentiation process and, upon further review, may afford a better understanding of the requirements for the conversion in terms of biochemical cues. The interactions between proteins in the pathway were also informative in that the connections between each pair were assigned a color and symbol to designate a particular relationship such as regulation or molecular transport. In addition, the visualization of the relative expression of each protein in the pathway provides additional insight into how these cells differentiate.

Using Pathway Assist to Integrate with Other Data Sources

We were easily able to import DIP, BIND, and the gene expression data. The expression levels of many of the genes in the pathway diagram were shown nicely. However, despite the easy load of DIP and BIND data, integration was fruitless. This is for two principle reasons: First, DIP uses a different nomenclature (SWISSPROT names) and set of accessions (Genbank, SwissProt, and PIR) than PathwayAssist does by default. Although PathwayAssist can certainly add SWISSPROT names and accessions, this needs to be done either manually, or by some other process. So even though a relevant interaction may be in the PathwayAssist database, it won't show up on the pathway diagram when automatically building pathways without special intervention. Second, most existing interaction data is specific to yeast. For mouse, very little interaction data currently exists in DIP or BIND. Thus even if the nomenclature issues were resolved in the software it would be of no avail until more mouse data is available.

CONCLUSIONS

Although the data that is currently available for the mouse is not as extensive as many simpler organisms, PathwayAssist was still able to provide a fair amount of information concerning the differentiation of murine stem cells into cochlear ear hair cells. The data that was available was primarily from ResNet, but as DIP and BIND continue to grow the amount of knowledge that can be gained through the use of PathwayAssist will be substantial. Unfortunately, even the graphical forms of these pathways can become enormously complex with a limited amount of data and one can conclude that as more data becomes available, a better way of organizing these pathways will be necessary.

The study by Rivolta et al along with other studies involving the use of gene microarrays has made it is apparent that this type of expression data is often of poor quality. Furthermore, due to irregular analysis schemes, the conclusions given in at least some scientific publications may be incomplete and unjustified. Possible solutions to this problem include refinements to the currently available microarrays, standardized analysis protocols, and new high-throughput methods for analyzing cellular activity.

Nevertheless, any additional information concerning the differentiation processes of stem cells will lead to advancements in tissue engineering and cell-based therapy. These relatively new methods of treatment rely heavily upon gaining a complete understanding of various cellular activities and their associated biochemical pathways. Since this can only be achieved through further experimentation and more elaborate data analysis tools, a new experimental protocol that is designed with PathwayAssist or similar tools in mind may be the key to describing the unknown pathways in their entirety.

APPENDIX

Table 1 – Gene Expression Fold Change Assessment

Note: An indication of “NA” means that due to missing data (negative average difference values), no fold change could be assessed.

Rivolta et al gene/protein name	Rivolta et al fold change assessment	Our fold change assessment
Cyclin A	Down	Confirmed
Cyclin B1	Down	There were 3 different features labeled Cyclin B1 on the MU11K chip set
Cyclin B2	Down	Confirmed
Cyclin E	Down	Confirmed
Cyclin F	Down	Confirmed
Gadd45	Up > 15 fold	There were 3 different features labeled Gadd45 on the MU11K chip set. The one likely referred to here had a “NA” on day 0, so the fold can’t be accurately assessed.
Chop-10	Up > 15 fold	Had a “NA” on day 0, so the fold can’t be accurately assessed.
Gas1	Up > 15 fold	Confirmed - Fold change varied, but it was generally high.
Decorin	Increased from day 2	Confirmed
Col6a1	Increased from day 2	Confirmed
Col3a1	Increased from day 2	NA – Had negative value at day 0.
Matrix Gla	Increased from day 2	Confirmed
OSF-2	Increased from day 2	NA – Had negative value at day 0
Col8a1	Detected at day 4	Detected from day 0
Col6a2	Detected at day 4	Detected from day 0
Laminin β 2	Detected at day 4	Detected from day 0
Laminin β 3	Detected at day 9	NA – Had negative value at day 0
CRBP1	Increased at day 2	Confirmed
IGFBP-2	Increased at day 2	There were 3 different features that may have been IGFBP-2 on the MU11K chip set
TGF β 3	Expressed at day 4	Expressed on all days
IGFBP-5	Expressed at day 9	Expressed on all days
C/EBP δ	Increased almost 10 fold at day 2 and remained high during late differentiation	NA – Had negative value at day 0
Prx2	Up regulated early in differentiation, but dropped below 5-fold by day 4	Confirmed - although numbers don’t match
Goosecoid	Peaked at day 4 with a 6-fold increase	Peaked at days 9 and 14

Notch1	Present throughout, reached peak after 3-4 days.	There were 5 different features labeled Notch1 on the MU11K chip set
Notch2	Not expressed	There were 2 different features labeled Notch1 on the MU11K chip set. Both were always expressed on each day.
Notch3	Detectable after day 2; reached peak at days 3-4.	Confirmed
Notch4	Not expressed	Always expressed on each day.
Hes1	Present at day 0, increased sharply after temperature switch, and dropped in days 3-4, constant till day 9.	NA – Had negative value at day 0. However, it apparently rose sharply at day 1.
Hes3	Steadily increased during differentiation and reached peak at day 11.	Reached peak at day 6, not day 11.
Hes5	Not detected	NA - Had negative value at day 0. However, it was detected on most other days.
Jagged2	Peaked at day 11	Confirmed

REFERENCES

1. Walgenbach, Klaus-J. et al. "Tissue engineering in plastic reconstructive surgery." *The Anatomical Record* 263 (2001): 372-376.
2. Tabata, Yasuhiko. "Recent progress in tissue engineering." *Drug Discovery Today* 6.1 (2001): 483-486.
3. Heath, Carole A. "Cells for tissue engineering." *Trends in Biotechnology* 18 (2000): 17-18.
4. Rivolta, M. et al. "Transcript profiling of functionally related groups of genes during conditional differentiation of a mammalian cochlear hair cell line." *Genome Research* 12 (2002): 1091-1099.
5. Li, Huawei et al. "Generation of hair cells by stepwise differentiation of embryonic stem cells." 100 (2003): 13495-13500.
6. Diehn, M. et al. "SOURCE: A unified genomic resource of functional annotations, ontologies, and gene expression data." *Nucleic Acids Research* 31 (2003): 219-223. URL: <http://genome-www5.stanford.edu/cgi-bin/source/sourceBatchSearch>.
7. Nikitin, et al. "Pathway studio – the analysis and navigation of molecular networks." *Bioinformatics* 19 (2003): 1-3.