國立交通大學

生物資訊研究所

碩士論文

利用時間序列之微陣列基因表現資料來比較人類和
老鼠在心臟胚胎發育的關係

Comparing Fetal Heart Development between Human and
Mouse based on Time-Series Gene Expression Profiles

研 究 生：任冠樺
指導教授：黃憲達 博士

中華民國九十六年六月

利用時間序列之微陣列基因表現資料來比較人類和老鼠在心臟胚胎發育的關係

Comparing Fetal Heart Development between Human and Mouse based on Time-Series Gene Expression Profiles

研究生：任冠樺　　Student：Kuan-Hua Jen

指導教授：黃憲達　　Advisor：Hsien-Da Huang

國立交通大學
生物資訊研究所
碩士論文

A Thesis
Submitted to Institute of Bioinformatics
College of the Biological Science & Technology
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Bioinformatics

June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

利用時間序列之微陣列基因表現資料來比較人類和老鼠在心臟胚胎發育的關係

學生：任冠樺　　　　　　　指導教授：黃憲達

國立交通大學 生物資訊研究所碩士班

中文摘要

心臟的發育是非常複雜的一個生理機制，有非常多的基因在心臟胚胎發育過程參與其細胞調控並決定了心臟的形成。微陣列（基因晶片）的實驗能夠一次大量產生許多基因表現的數據，在此研究中，想利用此一技術產生關於心臟發育時期的基因表現值。由於相關法令的問題，人類心臟胚胎的檢體獲取非常不易。為了更了解關於心臟發育的機制，故也利用老鼠心臟發育的胚胎來建立一個人類和老鼠的在心臟發育方面時間序列的平台。目前大部分的研究人員都使用一種物種來研究發育時期基因的變化，我們特別利用人類與老鼠心臟發育胚胎的時間序列檢體，並且使用兩物種間的同源基因和 dynamic time warping 演算法將此兩種物種作同源基因的分析，找出人類和老鼠中同源基因有相似變化的基因。而後再利用這些基因，做進一步的系統化分析，探討其功能和交互關係。此研究目的就是希望能利用基因表現的數據來更了解心臟發育過程中基因表現的模式與變化，並希望能發掘尚未被先前研究所探討的發育調控基因。

Comparing fetal heart development between human and mouse based on time-series gene expression profiles

Student：Kuan-Hua Jen　　　　Advisors：Dr. Hsien-Da Huang

Institute of Bioinformatics
National Chiao Tung University

Abstract

Heart development is a complex process involving many genes which control cell behavior in the embryo and determine its pattern, its form, and much of its behavior. Microarray experiments can generate an enormous amount of data at one time, so we use this technology to obtain gene expression profiles in heart embryonic development. But it is usually very difficult to obtain human heart fetus sample because of the issues of ethical, legal, and social consideration. In order to help us get more understanding of human heart development, we can use the mouse model system that is most often used. Therefore, we must establish a mapping system to make a cross bridge between these two species on developmental stages. To date, the vast majority of researches have focused their study on one species. Specially, we utilize orthologous genes and incorporate the dynamic time warping algorithm in order to map the time points that human and mouse gene expression profiles having highly correlated pattern. Firstly, we apply the algorithm to select the best time-warped orthologous genes having similar pattern. Then, these genes are clustered into groups. Each group has its unique mapping pattern and different biological meaning. The following task is to find relationship and pattern in distinct groups of genes, and to get close understanding into molecular process and gene function, mechanisms of embryogenesis of the heart, and comparative genomics. Ultimately, our aim is to achieve new insights into the heart developmental biology.

Acknowledgements

For my parents, and my dear friends.

Echo Jen 2007.7

Table of Contents

List of Figures

List of Tables

Chpater 1 Introduction

1.1 Affymetrix Gene Chip Microarray

In oligonucleotide microarrays (or single-channel microarrays), the probes are designed to match parts of the sequence of known or predicted mRNAs (see in **Figure 1.1**). There are commercially available designs that cover complete genomes from companies such as GE Healthcare[1], Affymetrix[2], Ocimum Biosolutions[3], or Agilent[4]. These microarrays give estimations of the absolute value of gene expression and therefore the comparison of two conditions requires the use of two separate microarrays. Oligonucleotide Arrays can be either produced by piezoelectric deposition with full length oligonucleotides or in-situ synthesis. Oligonucleotide Arrays are composed of 25-mer or 30-mer and are produced by photolithographic synthesis (Affymetrix) on a silica substrate or piezoelectric deposition (GE Healthcare) on an acrylamide matrix. Oligonucleotide microarrays often contain control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes.

[1] GE Healthcare: http://www.gehealthcare.com/worldwide.html
[2] Affymetrix: http://www.affymetrix.com/index.affx
[3] Ocimum Biosolutions: http://www.ocimumbio.com/web/default.asp
[4] Agilent: http://www.home.agilent.com/agilent/home.jspx?cc=US&lc=eng&cmpid=4533

There are a lot of researches that use the microarray technology to the study of mammalian organogenesis. It can provide great insights into the steps necessary to elicit a functionally competent tissue. Previous researches often focused on maybe one species embryo differentiation [1-3], sex determination of the mammalian gonad [4], gene expression patterns in one organ's development [5, 6], or analyzing expression profiles during the period from fertilization to implantation [7]. These studies that just mentioned never compare one organ between two species in embryonic development time. Our approach is to synchronize heart development stage between human and mouse and provide an opportunity to identify those functional genes that might be important for controlling embryogenesis and organogenesis.

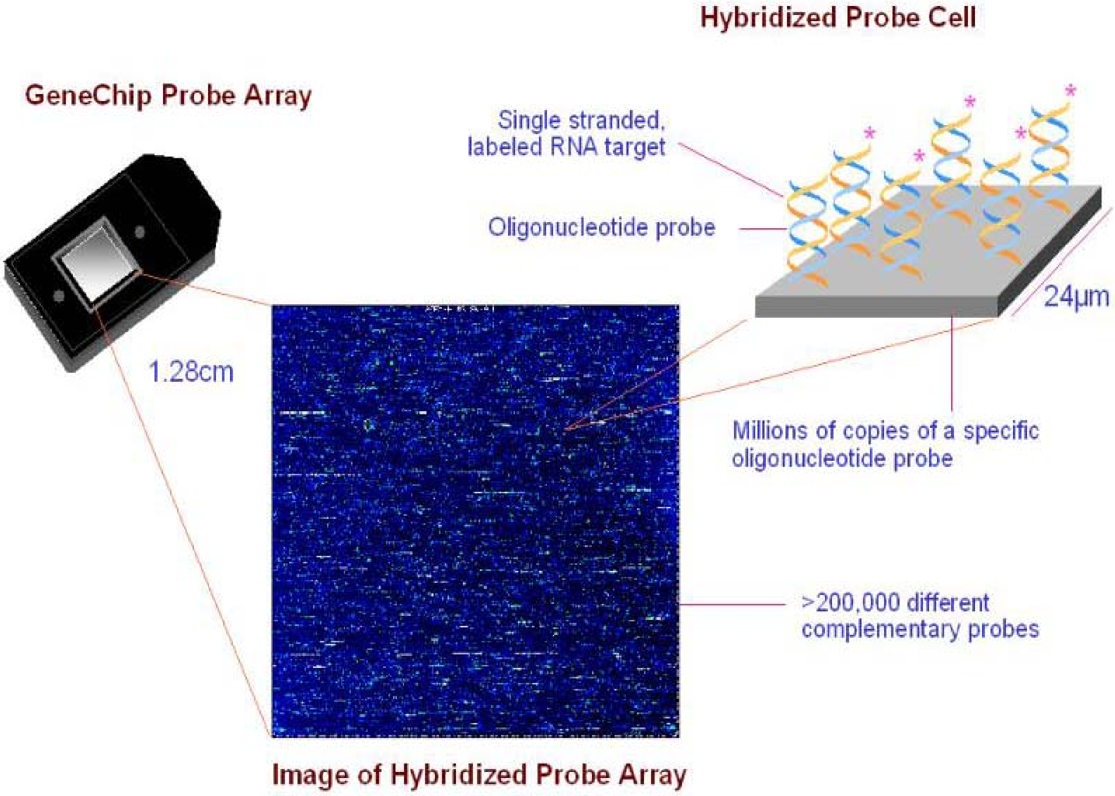

Hybridized Probe Cell
GeneChip Probe Array
Single stranded, labeled RNA target
Oligonucleotide probe
24µm
1.28cm
Millions of copies of a specific oligonucleotide probe
>200,000 different complementary probes
Image of Hybridized Probe Array

Figure 1.1 Affymetrix GeneChip Array.

1.2 Heart Development

The heart is the first organ to form during embryogenesis and its function is imperative and intricate from early on for the viability of the mammalian embryos. And it is the one of the few organs that has to function almost it is formed [8]. The developmental mechanisms that control the formation and morphogenesis of this organ have received much attention among classical and molecular embryologists. Due to the evolutionary conservation of many of these processes, major insights have been gained from the studies of vertebrate model. Heart development in all vertebrates follows the same general pattern: fusion of myocardium and endocardium in the ventral midline to form a simple tubular heart, onset of function, looping to the right side, chamber specification and formation, and at last, development of specialized conduction tissue, coronary circulation, innervation, and mature valves [9] (see in **Figure 1.2**).

Although, many genes important for heart development or organogenesis have been studied for a long time, global analysis of gene expression will provide more information about how the genes work and their interaction networks. In recent years, microarray technology has widely used for researchers to learn how genes' expression levels in different developmental stages, and to identify the cellular processes in which they participate.

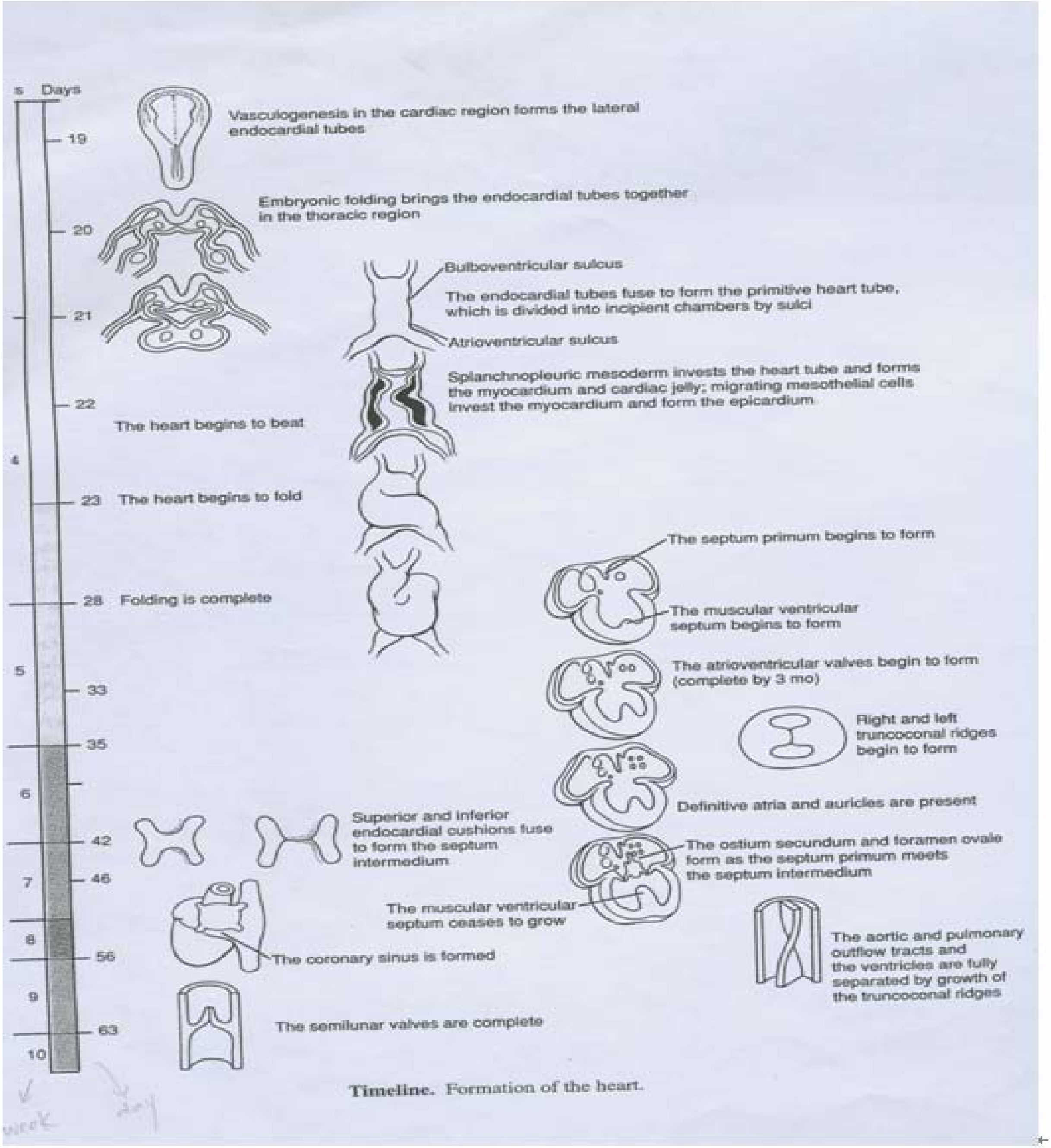

s Days
19
Vasculogenesis in the cardiac region forms the lateral endocardial tubes
20
Embryonic folding brings the endocardial tubes together in the thoracic region
21
Bulboventricular sulcus
The endocardial tubes fuse to form the primitive heart tube, which is divided into incipient chambers by sulci
Atrioventricular sulcus
22
Splanchnopleuric mesoderm invests the heart tube and forms the myocardium and cardiac jelly; migrating mesothelial cells invest the myocardium and form the epicardium
The heart begins to beat
4
23 The heart begins to fold
The septum primum begins to form
28 Folding is complete
The muscular ventricular septum begins to form
5
33
The atrioventricular valves begin to form (complete by 3 mo)
35
Right and left truncoconal ridges begin to form
6
Definitive atria and auricles are present
Superior and inferior endocardial cushions fuse to form the septum intermedium
42
The ostium secundum and foramen ovale form as the septum primum meets the septum intermedium
7
46
The muscular ventricular septum ceases to grow
8
56
The coronary sinus is formed
The aortic and pulmonary outflow tracts and the ventricles are fully separated by growth of the truncoconal ridges
9
63
The semilunar valves are complete
10
Timeline. Formation of the heart.
week
day

Figure 1.2 Formation of the heart.

1.3 Experimental Objectives

It is not practical to use multiple fetuses at the same gestational age to obtain statistical significances in gene expression level, because of the scarcity of useable fetal specimens at same gestational age. On the other hand, the change in gene expression along various fetal gestational weeks using the expression profiles derived from one fetus at a gestational age may be misleading, considering the existing variations among individual fetus even at the same age. Therefore, mouse has been adopted as a model system for studies of vertebrate development because of its similar features with human and favorable for genetic studies compared with other vertebrate systems. Using the mouse model will allow us to evaluate the changes in gene expression along various developmental stages, because we can use as many mice as necessary for each time point of a gestational age to eliminate the potential variations, which the result only from individual biological variations.

After mapping the gene expression profiles with the two species, we choose the best 250 match orthologous genes and cluster these genes into groups. As a preliminary analysis, each group of genes has its unique biological meaning after doing time warping. Moreover, specific characteristics were found to be associated with some features of the gene expression patterns. We employed an integrated analytical approach that encompasses Gene Ontology, biological pathway, and some previous research validations to provide more information for identifying the development-specific genes and get more understanding of their function in cardiogenesis. Our works presents a good example in which the combination of microarray technology with human and mouse model will not only consolidate our existing knowledge, but will also help us to identify novel factors that might be important for organogenesis. It also provides us with a global view on how genes are coordinated to form a genetic network to control heart embryogenesis.

The aims of this research are shown as below:

1. Constructing the mapping system between human and mouse
2. Aligning two different time series profiles by using microarray data
3. Identification of heart development-related genes
4. Understanding developmental related genes' function, pathway, regulation, and how they are coordinated to form a genetic network to control heart embryogenesis
5. Achieve new insights into the heart developmental biology

Chpater 2 Materials and Methods

2.1 Materials

2.1.1 Microarray Datasets

Affymetrix Human U133A and mouse 430A GeneChips have been successfully processed at the Genomic Medicine Research Core Laboratory (GMRCL) of Chang Gung Memorial Hospital.

Table 2.1 shows the detailed information of the dataset and platform of microarray data we used. There are no clearly defined development equivalences between the human and mouse fetus, in terms of gestational weeks for humans and post conception (p.c.) days or neonatal days (N) for mice. In this study we will analyze mouse at the following 16 time points (12, 13, 14, 15, 16, 17, 18, N1, N2, N3, N4, N5, N6, N7, N8, N9) and Human at 10 time points (6, 7, 8, 9, 12, 13, 16, 21, 23, 24). Almost each time point has performed one microarray experiment, but in mouse pc-14 day and pc-15 day, two replicates had done in this research.

Table 2.1 Microarray datasets using in the research.

Species	**Platform**	**Number of time points**	**Time points**	**Unit**
Human	Affymetrix Human U133A	10	6x1, 7x1, 8x1, 9x1, 12x1, 13x1, 16x1, 21x1, 23x1, 24x1	gestational weeks
Mouse	Affymetrix Mouse 430A	16	12x1, 13x1, 14x2, 15x2, 16x1, 17x1, 18x1, N1x1, N2x1, N3x1, N4x1, N5x1, N6x1, N7x1, N8x1, N9x1	post conception (p.c.) days and neonatal days (N)

2.1.2 Datasets from GEO Database

There are several public repositories for gene expression data, which, in time, are likely to serve a role for gene expression data similar to that of DDBJ/ EMBL/GenBank for sequence data. We found a dataset from Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) is also performed in the mouse heart embryonic development, and used it to validate our own data. The dataset in GEO is GDS627 (see in **Table 2.2**).

Table 2.2 Microarray datasets from GEO using in the research.

Species	Platform	Number of time points	Time points	Unit
Mouse	Affymetrix Mouse 430A (GDS627)	7	10.5x3, 11.5x3, 12.5x6, 13.5x6, 14.5x6, 16.5x6, 18.5x6	post conception (p.c.) days

2.2 Methods

2.2.1 Microarray Experiment

2.2.1.1 Collection of Human Specimens

Total RNA specimens of human heart from 6th to 12th week of gestational weeks were obtained from ViroGen Inc. (Watertown, MA, USA). Human abortuses of gestational weeks at 13, 16, 20, 21, 23 and 24 were donated by pregnant women with either cervical incompetence or premature preterm rupture of membrane, resulting in inevitable delivery of otherwise normal fetuses. Fetal organs were immediately kept in RNA*later* reagent (Ambion, TX. USA) at 4^{o}C for 24-48 h before transferred to –80^{o}C for long term storage. All pregnant women in this study signed an informed consent. This study was approved by the Internal Review Board (IRB) of Chang Gung Memorial Hospital.

2.2.1.2 Animal Experiment

Female C57/BL6 mice at 8 to 10 weeks of were used in this study. In the afternoon, four female mice were transferred to each cage containing one male mouse at 12 to 14 weeks old. On the next morning, each group of 4 female mice were transferred to a new cage and labeled as potentially post conception (PC) day 0. Pregnancy in female mice became visually detectable on PC day 10, and the fetuses were collected on the noon of PC day 12 through PC day 18. For this group of C57/BL6 mice, spontaneous delivery occurred on PC day 19, when the neonates were labeled as the neonate (N) day 0. Neonatal mice were collected from N day 1 to N day 9.

Collected through hysterectomy, fetal mice from PC day 12 to 15 were immediately immersed in RNA*later* at 4^{o}C for 48 h before organs were collected by dissection under microscopy. Mice at age from PC day 16 through N day 9 were sacrificed by cervical dislocation, and the dissected organs were immersed in RNA*later* at 4^{o}C for 48 h before RNA was extracted. Hearts from 4 to 8 fetal mice were pooled at each time point. The use of animal in this study had complied with the guidelines of Experimental Animal Committee and this study was approved by the Internal Review Board (IRB) of Chang Gung Memorial Hospital.

2.2.1.3 RNA Extraction and Microarray Analysis

The procedures of RNA extraction using TRIZOL (Invitrogen, Carlsbad, CA, USA) and RNAeasy purification kit (Qiagen Inc., Valencia, CA,USA), and confirmation of RNA quality and quantity with Agilent Bioanalyzer 2100 (CA, USA) were similar to previous reports [10-13]. Gene expression profiles in human fetal heart and murine heart were analyzed Affymetrix U133A GeneChip and 430A GeneChip, respectively.

2.2.2 Data Preprocessing

2.2.2.1 Normalization

There are a variety of reasons why the raw measurements of gene expression for two samples may not be directly comparable: the quantity of starting RNA may not be equal for each of the samples, there may be differences in labeling and detection efficiencies for the fluorescent labels, and there may be additional systematic effects that can skew the measured expression levels and the derived expression ratios. Normalization is any data transformation that adjusts for these effects and allows the data from two samples to be appropriately compared.

Robust Normalization accounts for probe set characteristics resulting from sequence-related factors, such as affinity of the probe set to the RNA and linearity of the hybridization of each probe pair. More specifically, this factor corrects for the inevitable error of using an average intensity of all the probes on the array as a normalization factor for every probe set. Robust Multi-array Analysis (RMA) was adopted due to its sensitivity and specificity in detecting differential expression and is a useful improvement to other kinds of normalization method for researchers using the GeneChip technology [14, 15]. The normalization results are presented in **Figure 2.1** and **Figure 2.2**.

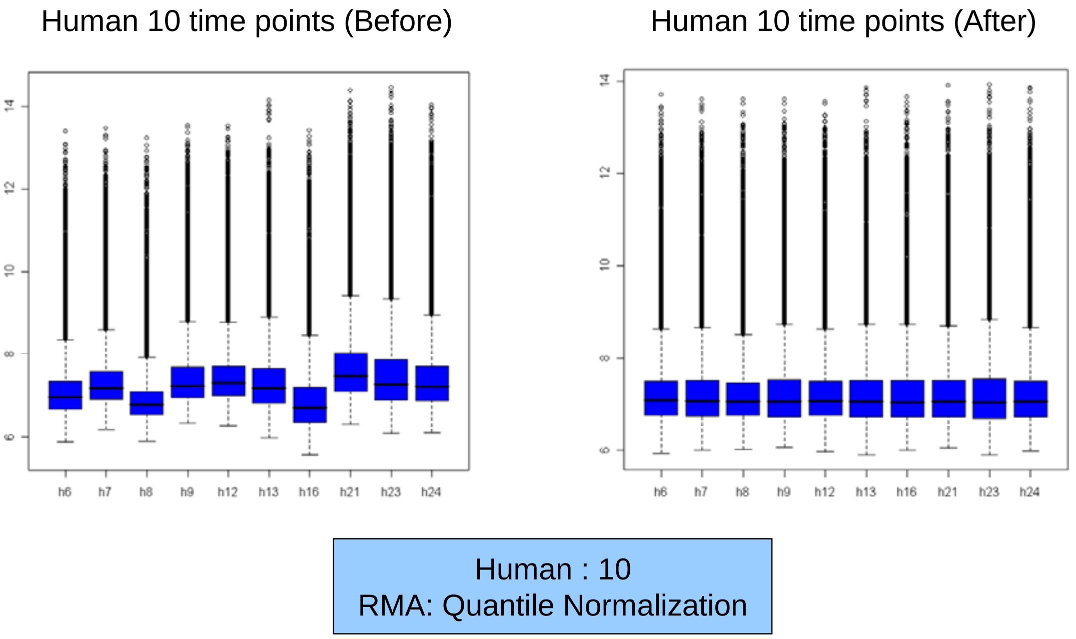


Figure 2.1 Normalization result of human data.

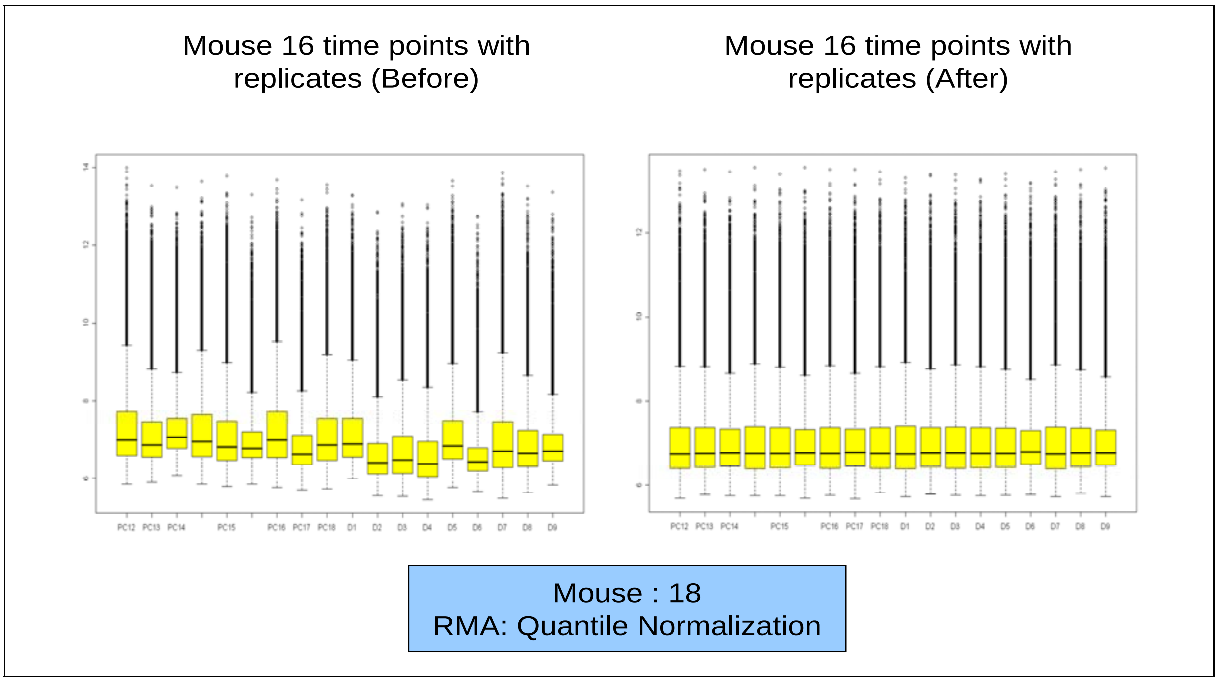


Figure 2.2 Normalization result of mouse data.

2.2.2.2 Use of replicate data

Replication is essential for identifying and reducing the effect of variability in any experimental assay, and microarray analysis is no exception. Biological replicated use independently derived RNA from distinct biological sources to provide an assessment of both the variability in the assay and the inherent biological variability in the system under study.

Biological replicates allow commonly expressed genes to be identified, as well as those that are distinct to the particular biological sample. In the research, we did average the replicated to produce a single consensus measurement and thereby reduce the complexity of the final data.

2.2.2.3 Data Filtering

The goal of most other transformations is to filter the dataset to reduce its complexity and increase its overall quality. Many are designed to flag questionable and low quality data, while others are used to identify differentially expressed genes or to enhance particular feature of the data. Below is our method. If more than one probe sets represented the same gene, their intensities were averaged. Then, all hybridization intensity values < 20, including negative intensity values, were raised to a value of 20, in order to prevent the too small and negative intensities in these datasets. If the continuous time-points expression profile of one single gene is too flat, we called it “smooth pattern”, that gene would be filtered out. We hope that each gene we use for the latter dynamic time warping algorithm has a specific expression pattern; it means that the gene has variable expression intensities at different developmental ages, and we guess maybe this gene control the embryogenesis and has an important role in heart development. We made the calculation for genes with all the time-point intensities smaller its mean ± 0.3*mean were excluded from the latter use of mapping. As a result, we collected only undulated genes with any intensity of variation of greater than mean ± 0.3*mean, and transformed the data to z-score. Finally, z-score values at transcriptome level were calculated to represent expression data of each gene.

2.2.2.4 Standardization

If a distribution is normal but not standard, we can convert a value to the Standard normal distribution table by first by finding how many standard deviations away the number is from the mean.

The number of standard deviations from the mean is called the z-score and can be found by the formula: $Z = \frac{x - \mu}{\sigma}$. Consider the gene expression matrices in **Table 2.3** and **Table 2.4**. They all represent the expression levels of genes G1-G9 for experimental conditions C1, C2, C3 and C4. **Table 2.3** is the original data and **Table 2.4** is the original data transformed into z-score (standardization).

Table 2.3 Gene expression data matrix Ⅰ.

Gene expression data matrix of absolute expression measurements after normalization for samples C1, C2, C3 and C4.

Gene	C1	C2	C3	C4	Mean	Std
G1	211.5703	168.1379	175.8446	180.5085	184.0153	19.06502
G2	199.3421	370.9393	450.259	413.8647	358.6013	111.0119
G3	292.1011	384.8857	330.9426	277.6322	321.3904	47.94283
G4	58.30043	57.17114	59.13815	57.66531	58.06876	0.849661
G5	289.157	362.7946	335.4638	346.5588	333.4935	31.61678
G6	126.1376	120.9111	140.856	126.5952	128.625	8.551966
G7	658.9924	686.8183	809.7875	701.4234	714.2554	66.07527
G8	46.54035	48.21487	51.91154	47.12361	48.44759	2.411336
G9	219.3456	253.1414	285.1363	243.8249	250.362	27.21356

Table 2.4 Gene expression data matrix Ⅱ.

Gene expression data matrix of expression measurements after standardization for samples C1, C2, C3 and C4

Gene	C1	C2	C3	C4
G1	1.445314	-0.8328	-0.42857	-0.18394
G2	-1.43461	0.111142	0.825657	0.497815
G3	-0.61092	1.324397	0.199241	-0.91272
G4	0.272665	-1.05644	1.258615	-0.47484
G5	-1.40231	0.926756	0.062316	0.413238
G6	-0.29085	-0.902	1.430197	-0.23734
G7	-0.83636	-0.41524	1.445807	-0.1942
G8	-0.79095	-0.09651	1.436528	-0.54907
G9	-1.13974	0.10213	1.27783	-0.24022

2.2.2.5 Identification of Orthologous Genes

Orthologs are genes that are related by direct evolutionary descent. The identification of orthologs is particularly important because these genes should play similar developmental or physiological roles, and consequently, their study in rodent or other models can provide insight into their functions in humans. We use orthologous genes to establish relations between human and mouse and then analysis their gene expression profiles with microarray data.

HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes. The genomes represented in the recent Build 52 of HomoloGene include *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, and so on [16]. This database contains 19157 orthologous genes between human and mouse.

Table 3.1 presents the preprocessing steps and detailed information of the microarray data we used. We have performed a novel bioinformatics study and use the orthologous genes to be the cross-bridge between human and mouse. At last, we concluded the number of orthologs (probe sets) included in U133A and 430A is around 15530. Therefore, we have a large set of common genes covered by both sets to do the comparative functional genomics study. **Figure 2.3** reveals the overview of our analysis of microarray data between human and mouse.

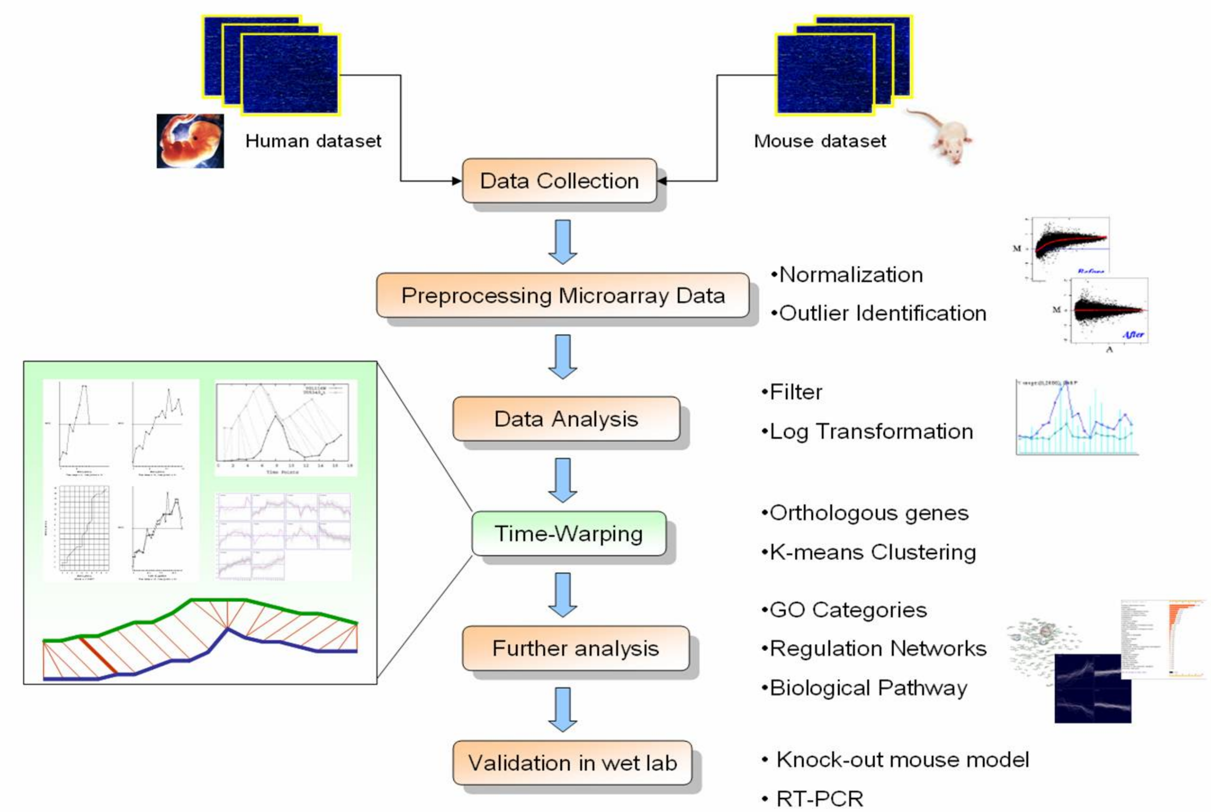


Figure 2.3 The overview of the microarray data analysis between human and mouse on the developmental stages.

2.2.3 Analysis of Gene Expression Data

The goal of microarray data analysis is to find relationships and patterns in the data and ultimately achieve new insights into the underlying biology. For instance, one could look for groups of genes having similar expression under similar conditions and try to find whether their products share similar functional roles in the cell, or for genes whose expression depends on the particular state of the system and see if the functions of their products can help to explain the particular phenotype.

2.2.3.1 Distance Similarity Measurements

Most of the gene expression data analysis methods are based on comparisons between the gene or sample expression profiles. In order to make these comparisons first we need a way to measure similarity or dissimilarity between these objects, i.e. between vectors representing genes or samples.

2.2.3.1.1 Euclidean Distance

Euclidean distance is the most common distance measure, and the one we use in everyday situations. Euclidean distance between points A = (a_1, a_2) and B = (b_1, b_2) in two dimensions can be expressed using Pythagoras's theorem:

$$D_{Eucl}\ (A,B) = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2}$$

In n-dimensional space for vectors A = (a_i,…,a_n) and B = (b_i,…,b_n), Euclidean distance can be expressed as :

$$D_{Eucl}\ (A,B) = \sqrt{(a_i-b_i)^2}$$

2.2.3.1.2 Pearson Correlation Distance

We assume that the arithmetic mean of each gene expression profile is zero. We will see that under this assumption the angle distance is closely related to the Pearson correlation coefficient. The two expression profiles A and B for four samples are given. These are represented by vectors in four-dimensional space: A = (a_1, a_2, a_3, a_4) and B = (b_1, b_2, b_3, b_4). We can calculate the mean value for each profile as:

$$\bar{a} = (a_1+a_2+a_3+a_4)/4 \ \text{ and } \ \bar{b} = (b_1+b_2+b_3+b_4)/4$$

And shift each profile "down" by its mean, i.e. obtain new vectors:

$$A^0 = (a_1 - \bar{a},\ a_2 - \bar{a},\ a_3 - \bar{a},\ a_4 - \bar{a}) \text{ and } B^0 = (b_1 - \bar{b},\ b_2 - \bar{b},\ b_3 - \bar{b},\ b_4 - \bar{b})$$

Their dot product equals:

$$A^0 \cdot B^0 = (a_1 - \bar{a})(b_1 - \bar{b}) + (a_2 - \bar{a})(b_2 - \bar{b}) + (a_3 - \bar{a})(b_3 - \bar{b}) + (a_4 - \bar{a})(b_4 - \bar{b})$$

In general, in n-dimensional space:

$$A^0 \cdot B^0 = \sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b})$$

If we divide this by n-1, we obtain the well=known expression for covariance, which is used to establish the degree of association between two or more distributions. Covariance is calculated in the same way as variance, except that there are multiple distributions. The variance can be thought of as a measure of the distance from the mean, or the “spread” of the data. Covariance is the generalization of variance for two distributions and can be expressed as:

$$\mathrm{Cov}(A, B) = \frac{A^0 \cdot B^0}{(n - 1)}$$

The normalized covariance gives the expression for linear correlation, also known as the Pearson correlation coefficient (PCC):

$$\mathrm{Cor}(A, B) = \frac{A^0 \cdot B^0}{|A^0|\,|B^0|}$$

In this way we see that the PCC between vectors A and B is the same as the angle distance between these vectors in normalized and mean centered space. For unrelated distributions the PCC is near 1 for a strong correlation and near zero for a weak correlation.

2.2.3.2 Clustering

The goal of gene expression data clustering is to group together genes or samples that have similar expression profiles. Clustering is currently the most popular method of gene expression matrix analysis. It can be useful for discovering “type” of behavior, for reducing the dimensionality of the data (allowing tens of thousands of genes to be represented by a few groups each containing genes that behave similarly), as well as for the detection of outliers in the data. Clustering is one of the unsupervised approaches to data analysis, which can be used in the absence of a priori information, or when annotations are not considered in the analysis.

2.2.3.2.1 Hierarchical Clustering

Hierarchical agglomerative clustering is a process in which the data are successively fused, typically until all the data points are included. For hierarchical agglomerative clustering usually all the pair-wise distances between objectives need to be defined. An agglomerative process typically starts by considering each object/data point as a separate, or singleton, cluster. Starting with n objects, the result of the first iteration of clustering is that the two objects that are most similar are grouped together to form a single cluster, leaving (n-1) clusters. The distance between the objects and the newly formed cluster containing two objects is then updated and the next most similar objects and clusters are grouped together as a single cluster[17]. The results of hierarchical clustering are frequently represented in a hierarchical tree, also known as a dendrogram (see in **Figure 2.4**).

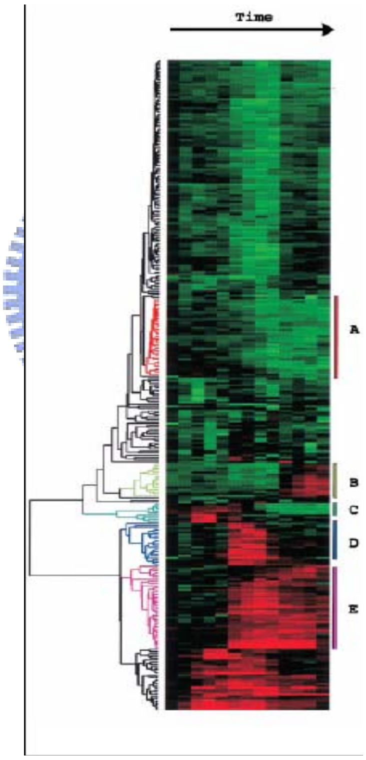


Figure 2.4 Hierarchical tree.

2.2.3.2.2 K-means Clustering

K-means is the most common method of partition-based clustering. It starts with the given number of cluster centers, chosen either randomly or by applying some heuristics. Next the distance from the centroids to every object is calculated, and each object is assigned to the cluster defined by the closest centroid; then, for each cluster the new centroid is found. The distance from each object to each of the new centroids is calculated and in this way the boundaries of the partitioning are revised. This is repeated either until the centroids stabilize or until an a priori defined maximum number of iterations has been reached (see in **Figure 2.5**).

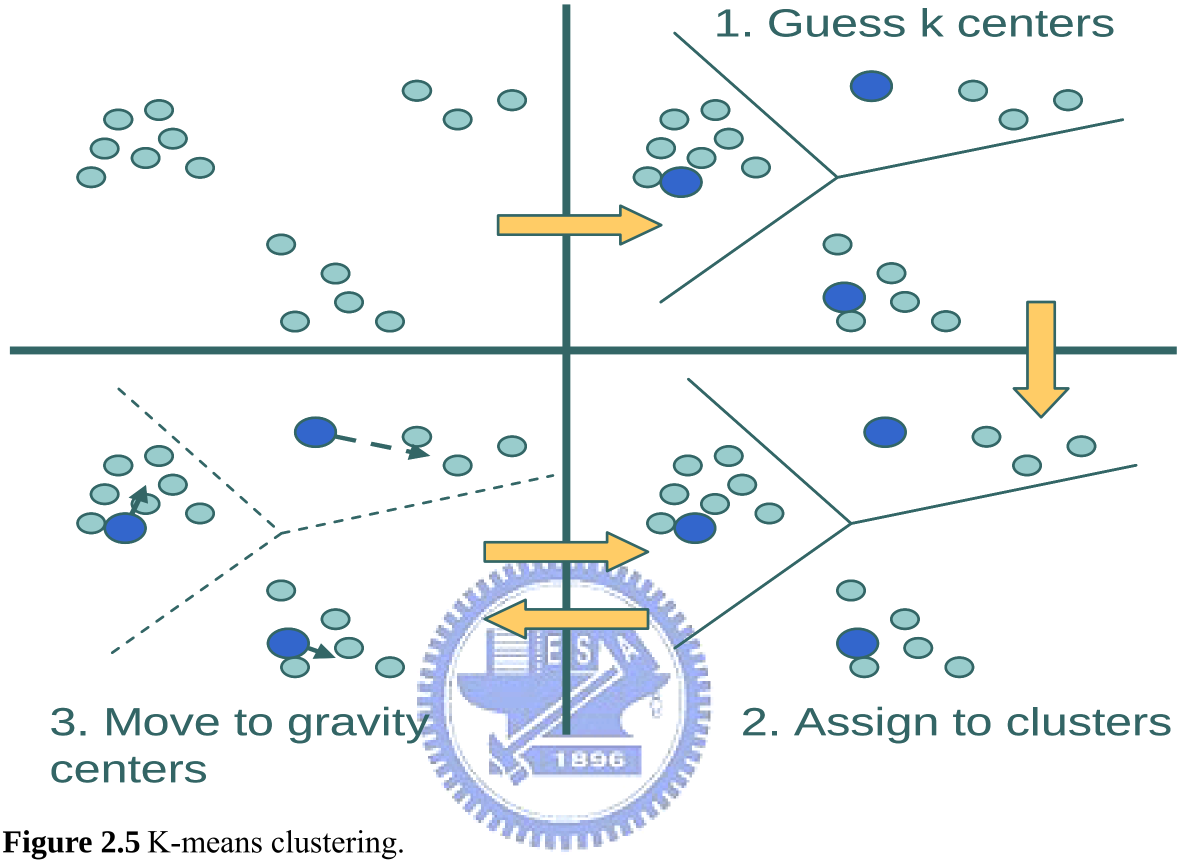
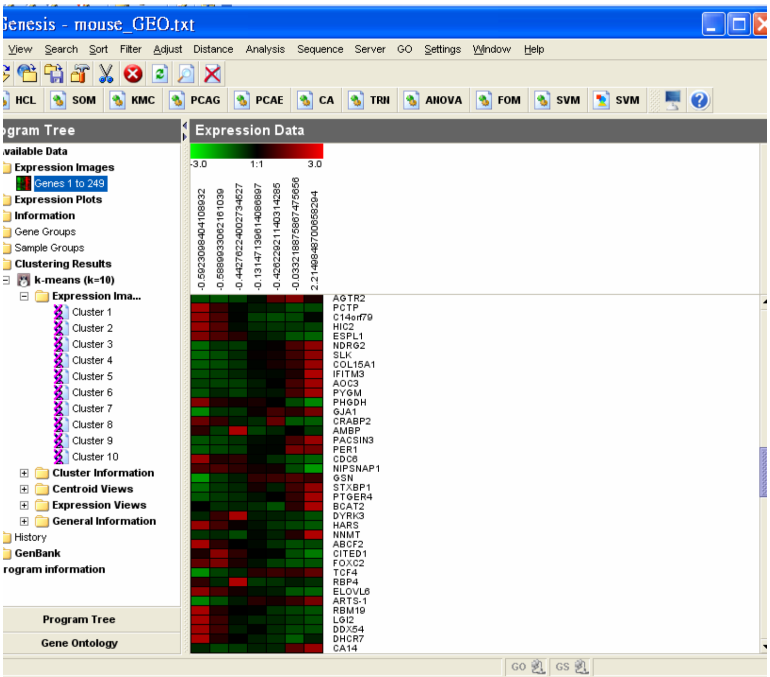


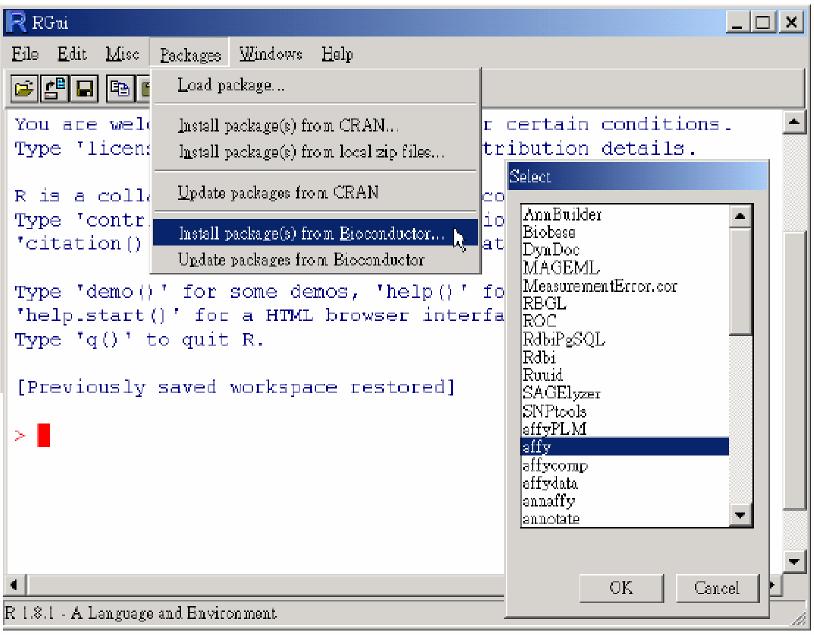
Figure 2.5 K-means clustering.

2.2.3.3 Software and Tool

2.2.3.3.1 Genesis

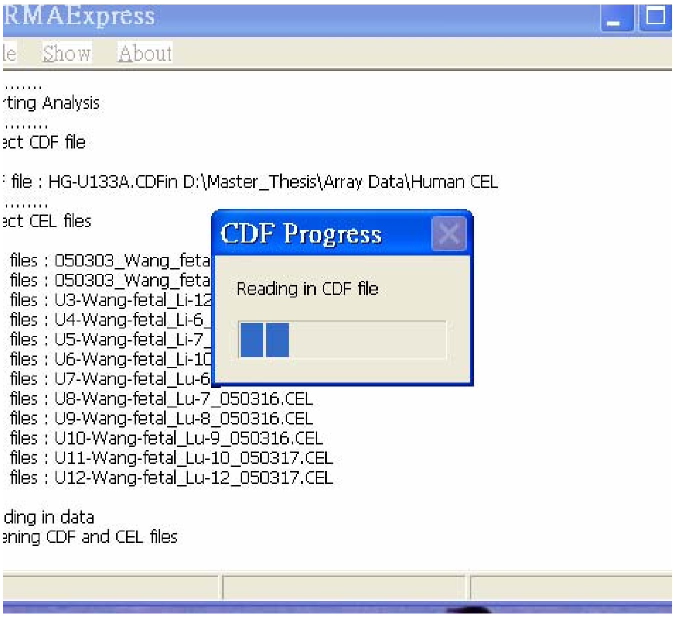


2.2.3.3.2 R


RGui
File Edit Misc Packages Windows Help
Load package...
Install package(s) from CRAN...
Install package(s) from local zip files...
Update packages from CRAN
Install package(s) from Bioconductor...
Update packages from Bioconductor
Select
AnnBuilder
Biobase
DynDoc
MAGEML
MeasurementError.cor
RBGL
ROC
RdbiPgSQL
Rdbi
Ruuid
SAGElyzer
SNPtools
affyPLM
affy
affycomp
affydata
annaffy
annotate
OK
Cancel
R 1.8.1 - A Language and Environment

```
You are wel                     r certain conditions.
Type 'licens                    tribution details.

R is a coll                     co
Type 'contr                     io
'citation()                     at

Type 'demo()' for some demos, 'help()' fo
'help.start()' for a HTML browser interfa
Type 'q()' to quit R.

[Previously saved workspace restored]

> 
```

2.2.3.3.3 RMAExpress

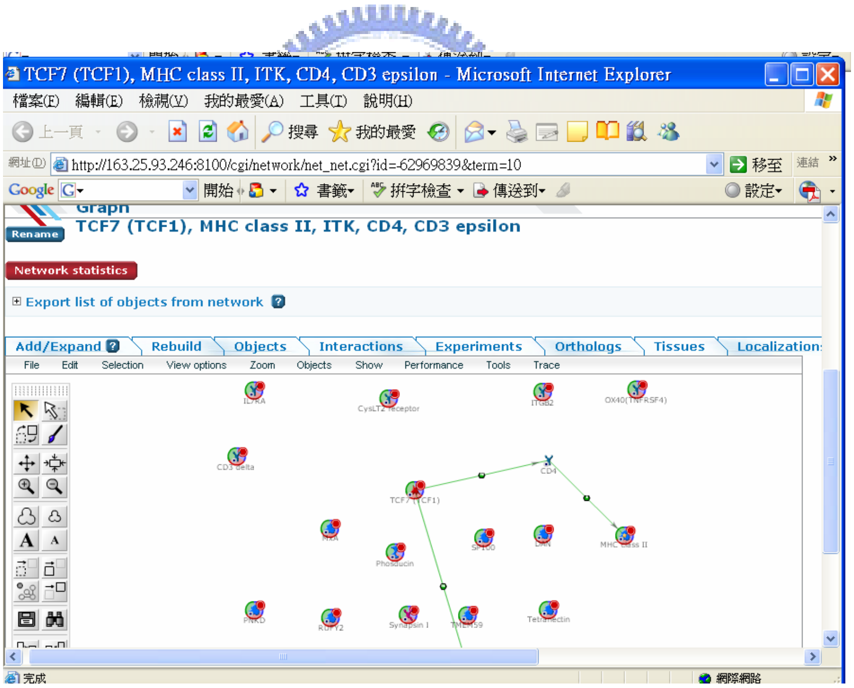


```
.......
ting Analysis
........
ect CDF file
 file : HG-U133A.CDFin D:\Master_Thesis\Array Data\Human CEL
........
ect CEL files

 files : 050303_Wang_feta
 files : 050303_Wang_feta
 files : U3-Wang-fetal_Li-12
 files : U4-Wang-fetal_Li-6_
 files : U5-Wang-fetal_Li-7_
 files : U6-Wang-fetal_Li-10
 files : U7-Wang-fetal_Lu-6
 files : U8-Wang-fetal_Lu-7_050316.CEL
 files : U9-Wang-fetal_Lu-8_050316.CEL
 files : U10-Wang-fetal_Lu-9_050316.CEL
 files : U11-Wang-fetal_Lu-10_050317.CEL
 files : U12-Wang-fetal_Lu-12_050317.CEL

ding in data
ening CDF and CEL files
```

CDF Progress

Reading in CDF file

2.2.3.3.4 MetaCore



2.2.4 Time Series Data Analysis

Time series experiments provide a particular type of gene expression profile, revealing information about the order and the time scale of the expression events. In our research, we wish to compare gene expression time series data from different experiments corresponding to two similar species. An example of such an approach is comparing gene expression during the cell cycle for cell cultures synchronized using time warping [18]. If some of the genes involved in the process under study are known, we can "synchronize" the periods, by comparing the expression level of these known genes.

Suppose the gene expression profiles of these two species is subject to variation, so that a function may be traced out more slowly during one portion and more quickly during another, and suppose these variations differ from one occasion to another. To allow for such variations when comparing functions, it is necessary to distort or "warp" the time axis appropriately, i.e., compressing it at some places and expanding it at others. The process of inferring the necessary compressions and expansions is often called time-warping [19]. **Figure 2.6** demonstrates the concept of time warping.

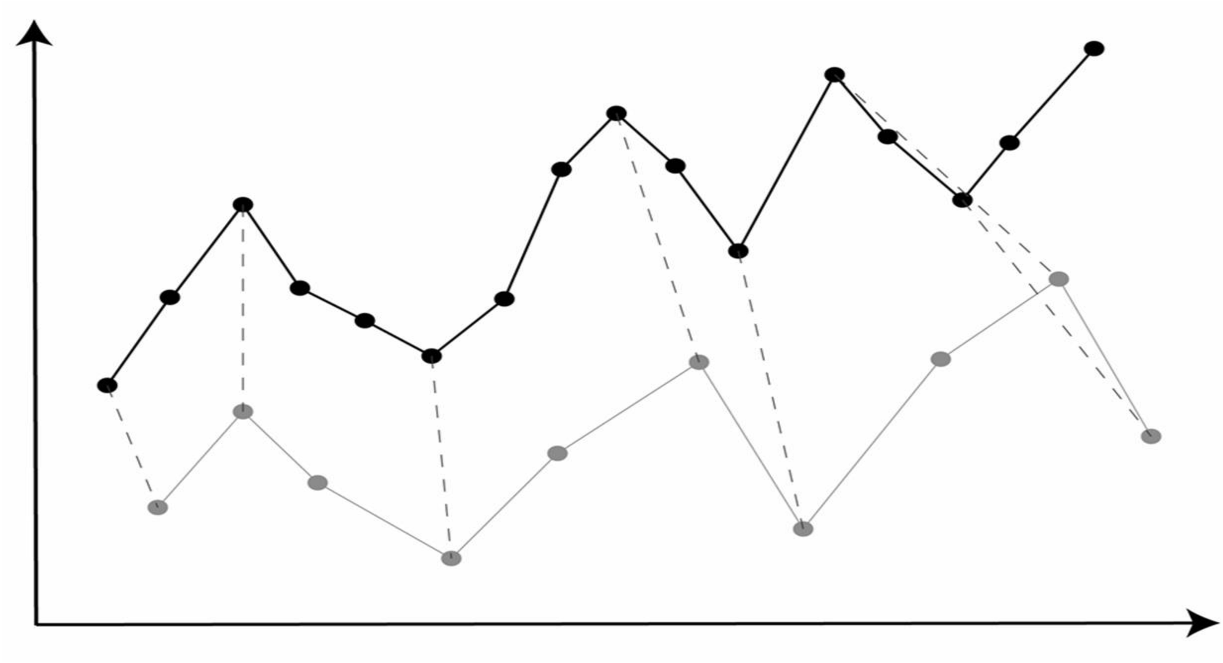


Figure 2.6 The concept of time warping.
The two time series have different rate of their expression level. In general, we need to use a distance measure, for which the time points of one series can match to the other.

2.2.4.1 The Concept of Time-Warping

Biological processes have the property that multiple instances of a single process may unfold at different and possibly non-uniform rates in different organisms, strains individuals, or conditions. For instances, different individuals affected by a common disease may progress at different and varying rates. This presents an issue for analysis of biological processes using time series of RNA expression levels: To find the time point of one series that corresponds best to that of another, it is insufficient to simply pair off points taken at equal measurement times. Analysis of such time series may therefore benefit from the use of alignment procedures that map corresponding time points in different series to one another.

An important area of application of these techniques is the study of biological processes that develop over time by collecting RNA expression data at selected time points and analyzing them to identify distinct cycles or waves of expression.

2.2.4.2 Time-Warping Programs

In our research, we have the datasets of the same biological condition which are human and mouse heart on developmental stages. In order to compare these two heart developmental time series in different species, we apply two time warping programs genewarp and grphwarp [18]. genewarp performs a simple time warping and grphwarp is a graphics generation program that take a file produced by genewarp.

While genewarp can be used on any set of genes regardless of whether their individual time course expression profiles are similar, we first applied it to all orthologous genes respectively so that they could be aligned. But these orthologous genes maybe don't have similar profiles; we have to choose the best 250 "mapping genes". These "250 mapping genes" have two characteristics. Firstly, they are all orthologous gene pairs. Secondly, each pair of them have similar expression pattern after doing "time-warping". It had been known that genes maybe have different expression patterns during the same biological process. We cluster these 250 genes into distinct groups according to their expression profiles. Therefore, genes in the same cluster have similar expression pattern and maybe the same biological function.

Chpater 3 Results

3.1 Large-scale transcriptional analysis of the developing heart

Approaches using DNA microarray have been successful in studying genome-wild transcriptional regulation during animal development, but suffer from several limitations. On multicellular organisms, cell division and differentiation leads to an increase in tissue complexity throughout development, but whole-animal microarray analysis cannot document this spatial information. We attempt to isolate mRNA form single tissue (Heart) at different developmental stages, measure gene expression, and assign expression to every gene at every time, in order to recreate the entire developmental expression pattern. Affymetrix oligonucleotide microarray platform has been used worldwide more than 1618 reports compiled in the NCBI PubMed Medline, till Jun 2007. The Affymetrix GeneChip system has been requires user to follow a strict manufacture's protocol. Therefore, Affymetrix system has been considered as a relatively stable platform and proved to be acceptable by the worldwide research community. According to its consistency and comparability of Affymetrix platform, we use U133A for human and 430A for mouse to do this research.

Affymetrix Human U133A and mouse 430A GeneChips have been successfully processed at the Genomic Medicine Research Core Laboratory (GMRCL) of Chang Gung Memorial Hospital. Furthermore, we have performed a pilot bioinformatics study and concluded the number of orthologous gene (transform to gene symbol ID) included in these two kinds of commercial GeneChips is around 8578. These orthologous genes are very prominent material to establish a cross-bridge between Human and Mouse. Therefore, we have a large set of orthologous genes covered by both chips to do the comparative functional genomics study.

After preprocessing the array data, there has 3490 orthologous genes between human and mouse chip (see in **Table 3.1** (b)). We used these genes for further analysis.

Table 3.1 Preprocessing of the microarray data and the number of genes after many steps of processing.

Gene Chip	Human Genome U133A	Mouse Genome 430A	Description

Probe sets	22283	22690	Number of probe sets on the chip
Total genes	13477	14218	Probe set ID transform to gene symbol
Orthologous genes	8578 (a)		Overlapped orthologous genes between human and mouse
Non-smooth genes	7919	7934	Filtering flat expression genes
Orthologous genes	3490 (b)		Overlapped orthologous genes between human and mouse
Time Warping	3490		Single orthologous gene pair time warping
Time Warping	250		Select 250 genes which distance scores are the least

3.2 Construction the Mapping System of Human and Mouse Microarrays

There are no clearly defined development equivalences between the human and mouse fetus, in terms of gestational weeks for humans and post conception (p.c.) days or neonatal days (N) for mice. Thus, in this study we will analyze mouse at the following 16 time points (12, 13, 14, 15, 16, 17, 18, N1, N2, N3, N4, N5, N6, N7, N8, N9) and Human at 10 time points (6, 7, 8, 9, 12, 13, 16, 21, 23, 24). **Table 2.1** shows the detailed information of the dataset and platform of microarray data we used. We use computational methods to provide a novel approach utilizing the gene expression profiles to match these two species with orthologous genes and select the best matching time-points which the expression patterns are highly correlated. Results from this study we propose here will provide the first-in-the-world complete data, at the transcriptome level, about the fetal development equivalence between the human and mouse.

3.2.1 Time-Warping for the Orthologous Genes

Orthologs are genes in different species that have evolved from a common ancestral gene by speciation and generally retain a similar function in the course of evolution. When mapping the expression profiles of human and mouse, using orthologous genes is a good way. In this approach, we use orthologous genes covered by human and mouse affymetrix microarray platform. We do the time-warping for each pair of orthologous gene in order to find their similarity of time series expression data. Time warping considers the similarity of pairs of vectors (orthologous gene) taken from a common k-dimensional space (feature space) taken one from each time series. **Figure 3.1** illustrates the time warping result of one orthologous gene between human and mouse.

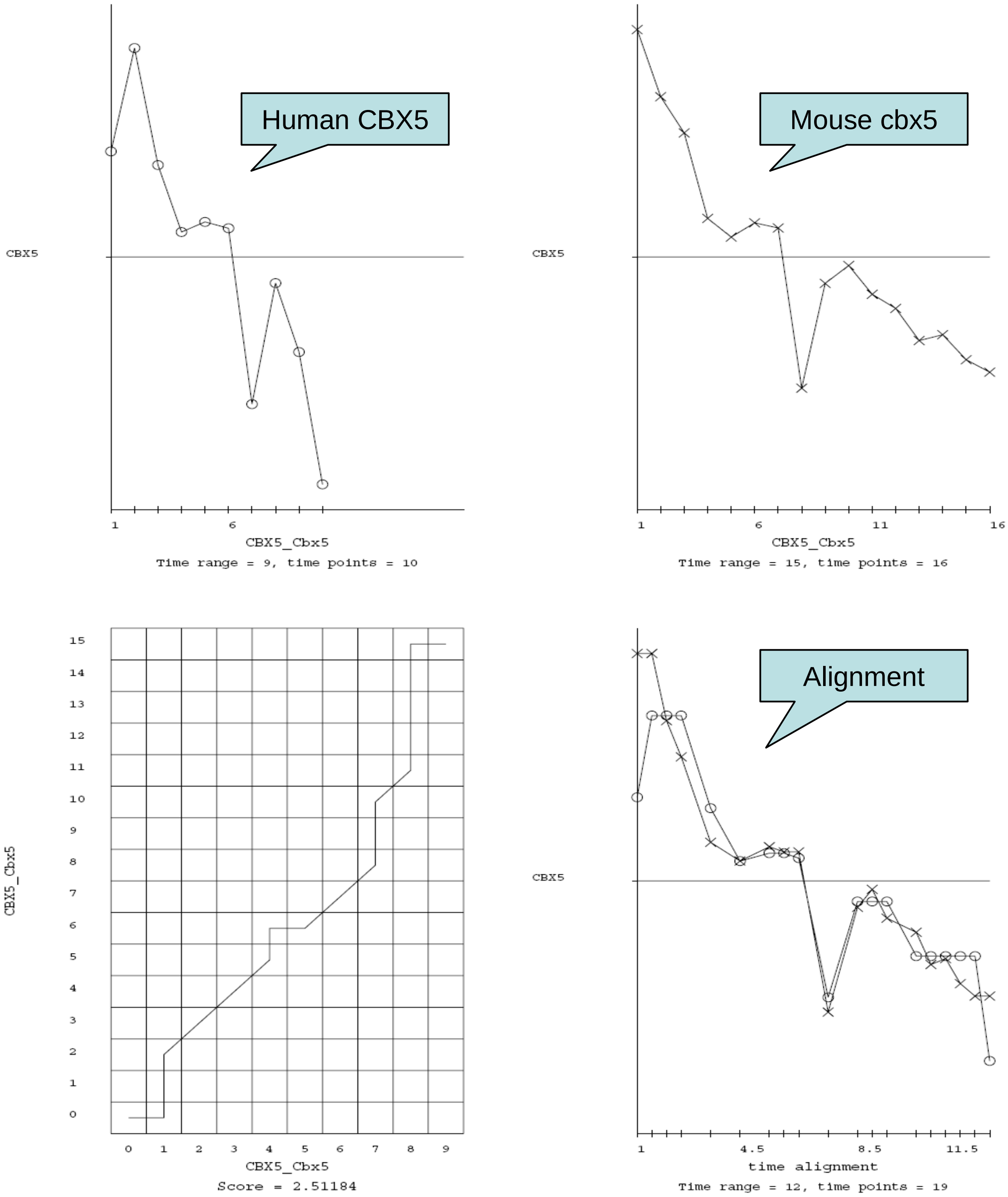


Figure 3.1 Time warping results of CBX5 and cbx5.

CBX5 is a chromobox homolog 5 gene, and its orthologous gene in mouse is cbx5. We got their gene expression profiles by the order of developmental time points. Top-left is the gene expression values of CBX5 in 10 time points; Top-right is the gene expression values of cbx5 in 16 time points. After applying the dynamic time warping program, genewarp, their expression profiles can map to each other like global alignment. Bottom-left is an alignment grid for CBX5 and cbx5. Every alignment corresponds to a path in the alignment grid from (0, 0) to (n, m). The entire alignment is simply a path (0, 0) →(1, 0) →(1, 1) →(1, 2) →(2, 3) →(3, 4) →(4, 5) →(4, 6) →(5, 6) →(6, 7) →(7, 8) →(7, 9) →(7,10) →(8, 11) →(8, 12) →(8, 13) →(8, 14) →(8, 15) →(9, 15) from (0, 0) to (n, m) in the grid. The score means the similarity of these two time series, the lower the score evaluates; the more similar the two time series are. In this case, the score of these two profiles is 2.51184. Bottom-right is the time alignment of these two profiles.

3.2.2 K-means Clustering of Time-Warped Genes into Distinct Groups

After "warping" for each single orthologous gene, we choose the best similar 250 genes for further analysis. **Figure 3.2** demonstrates the distance scores of 3490 othologous gene from minimum to maximum. The fewer the distance score, the more similar the orthologous gene pairs. **Figure 3.3** displays the distribution of 3490 orthologous genes. The selected 250 gene pairs have very similar expression pattern after time warping. It means that these genes are co-expressed in human and mouse heart development. These genes have three characteristics: (1) They are all orthologous gene pairs. (2) They are developmental related genes, especially expressed on embryonic stages. (3) After time warping, they have similar expression patterns. We called that "co-expressed gene pairs". Each "co-expressed gene pairs" have its unique matching time points. In order to know these genes more systematic, we use k-means clustering to divide these 250 genes into 12 groups. In order to make our data more authentic, we combined mouse and human expression data to do the clustering in order to make the human and mouse data more correspondent with each other. After clustering, each group of genes has very similar pattern. **Figure 3.4** shows k-means clustering result of the 250 orthologous genes. The annotations of the 250 genes are listed in Appendix A.

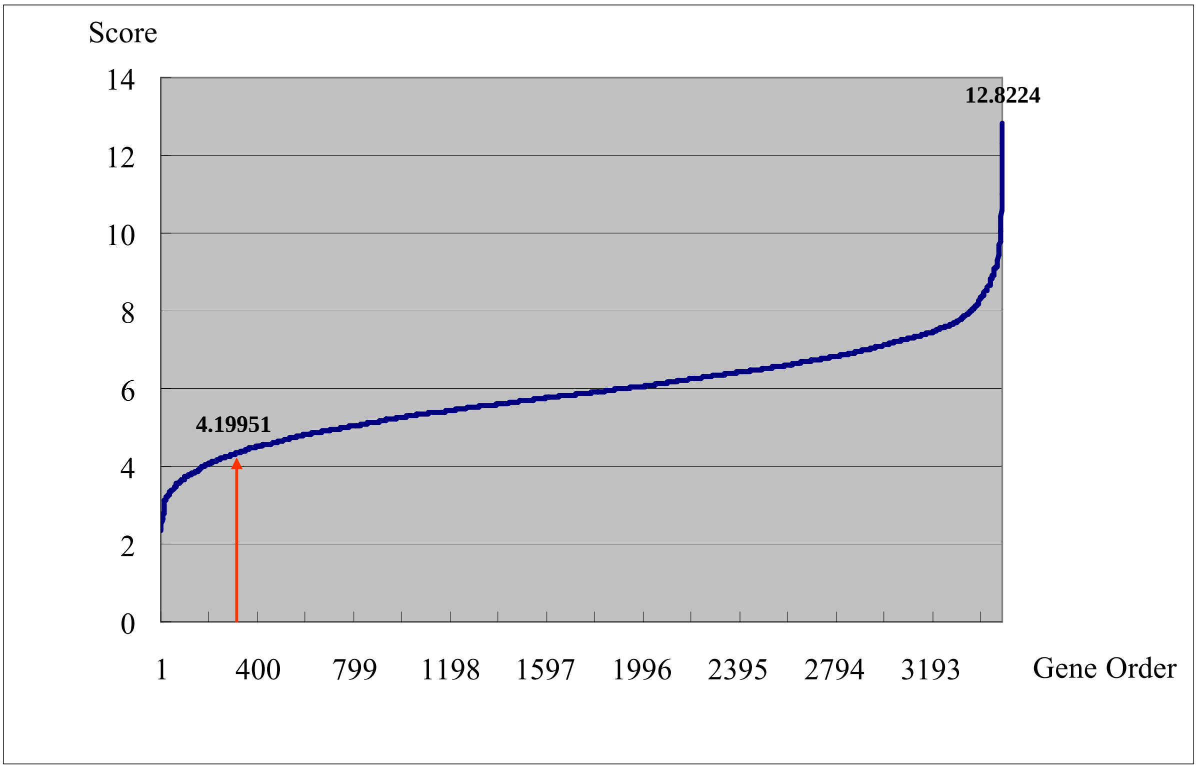


Figure 3.2 The distance score of 3490 orthologous genes pairs from the minimum to the maximum.

The minimum score of gene (FZD2) is 2.35004; and the maximum score of gene (FLJ10847) is 12.8224. We select the most similar 250 genes which distance scores are less than about 4.2, and use these genes for further analysis.

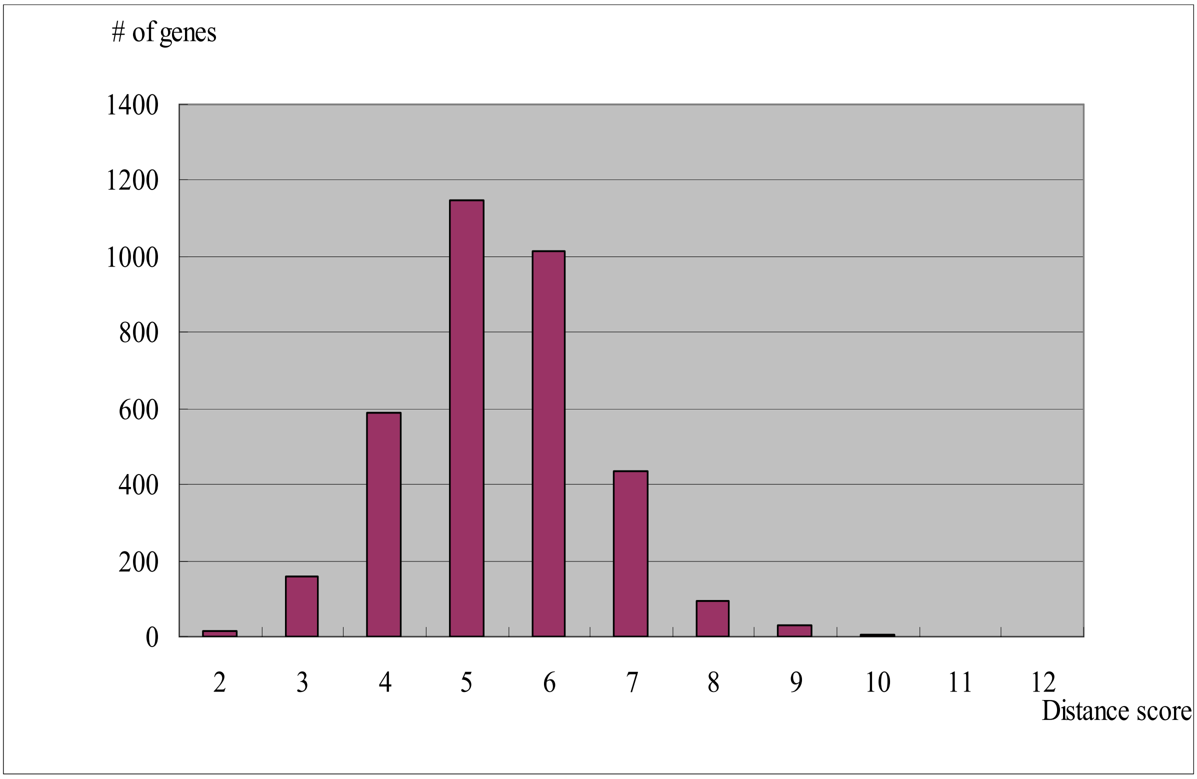
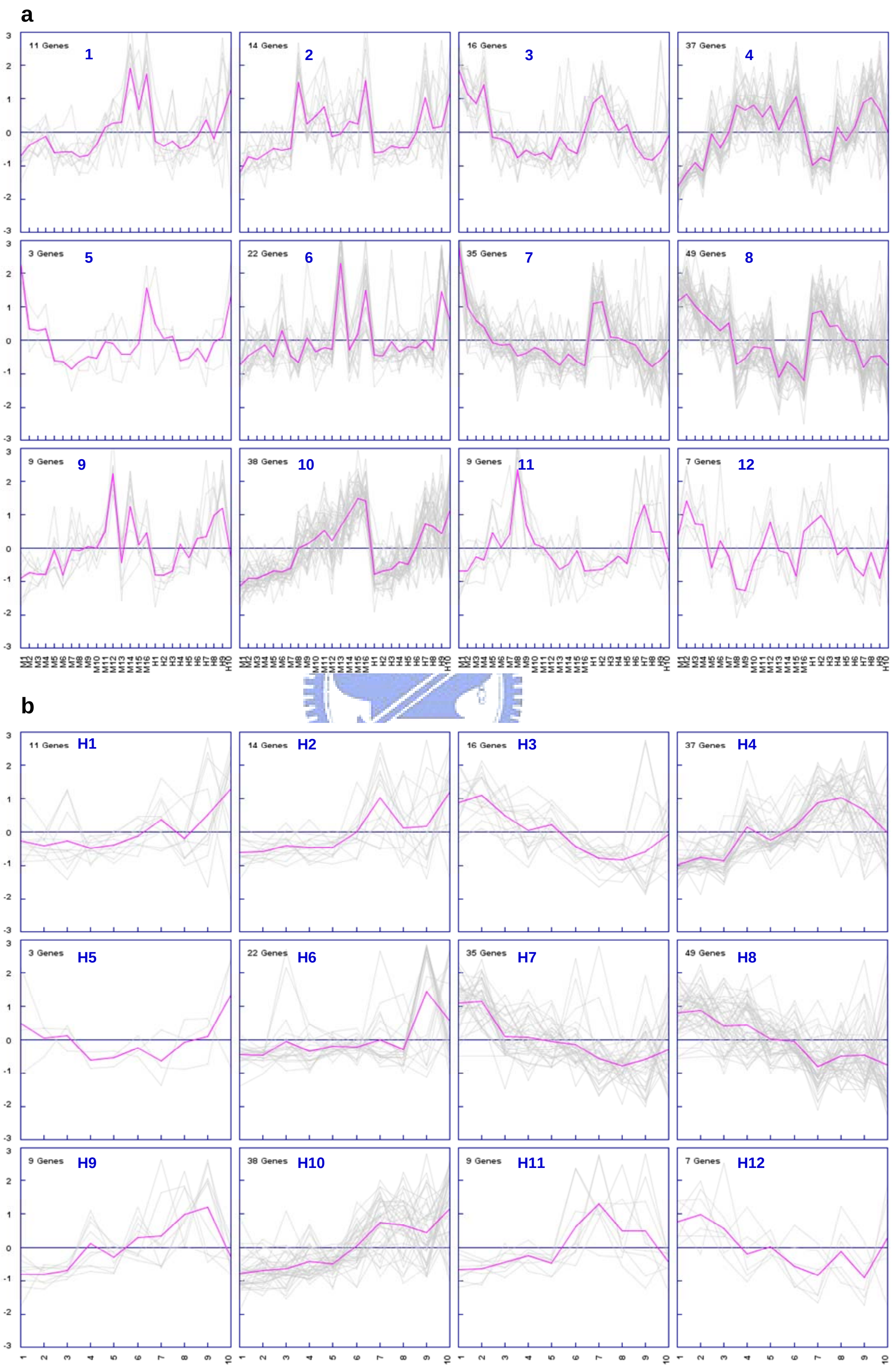


Figure 3.3 The distribution of distance scores of the 3490 orthologous gene pairs.

Most genes' scores are less than 6 and more than 5. There are just two gene's distance scores more than 12.


a
11 Genes 1
14 Genes 2
16 Genes 3
37 Genes 4
3 Genes 5
22 Genes 6
35 Genes 7
49 Genes 8
9 Genes 9
38 Genes 10
9 Genes 11
7 Genes 12
3
2
1
0
-1
-2
-3
M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 H1 H2 H3 H4 H5 H6 H7 H8 H9 H10
b
11 Genes H1
14 Genes H2
16 Genes H3
37 Genes H4
3 Genes H5
22 Genes H6
35 Genes H7
49 Genes H8
9 Genes H9
38 Genes H10
9 Genes H11
7 Genes H12
1 2 3 4 5 6 7 8 9 10

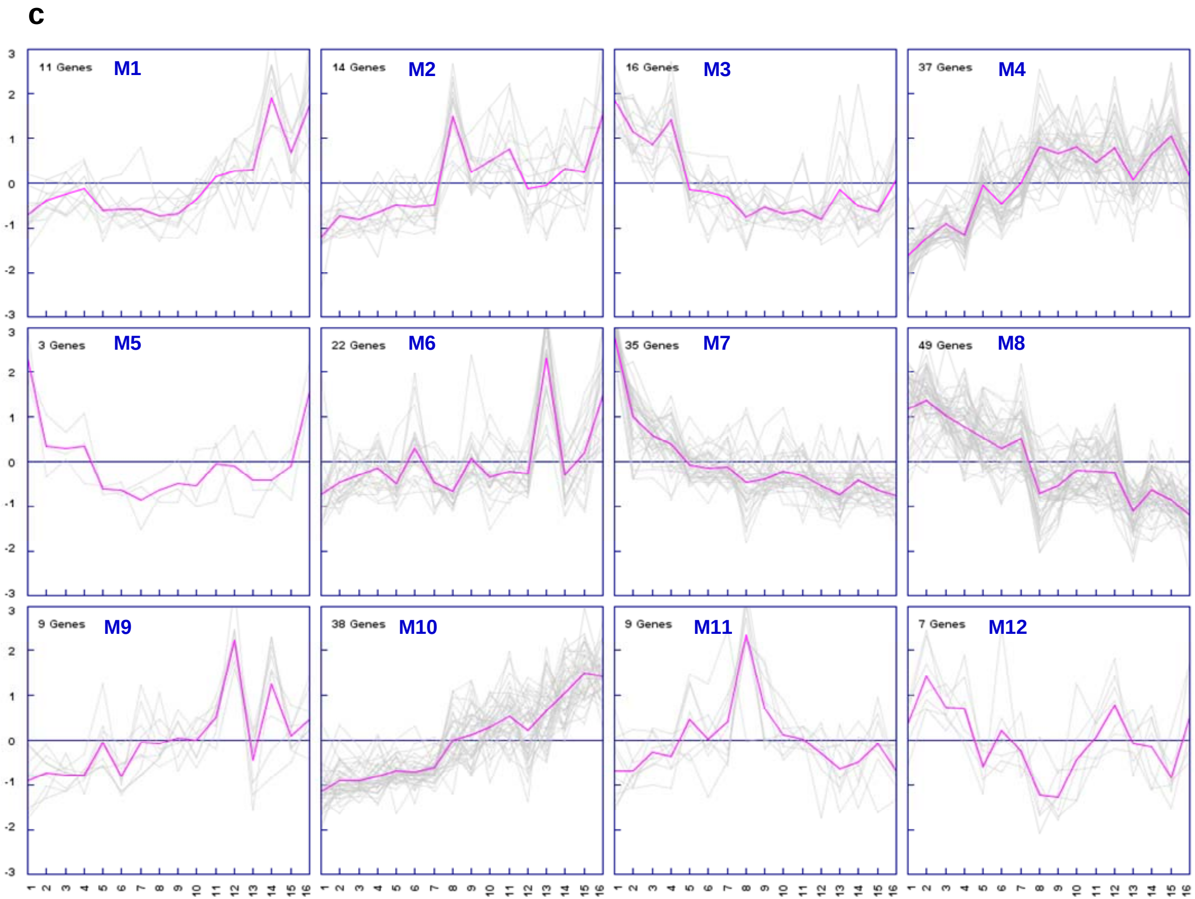


Figure 3.4 K-means clustering of Human and Mouse 250 genes.

(a) K-means clustering of combining human and mouse 250 genes expression values. First 16 time points are the mouse values; Last 10 time points are the human values. Each box illustrates the expression values (log2 ratio) of genes in this group and how many genes clustering into this group. (K=12 and the distance measurement is Pearson correlation coefficient).

(b) The 12 groups of human 250 genes.

(c) The 12 groups of mouse 250 genes correspondent to their human orthologous genes. Each group of genes has the similar expression trend with their correspondent human group so it is very appropriate for the next step---time-warping within the same group of genes between human and mouse.

3.3 Time Warping for Each Cluster

Each cluster contains many genes, which have similar expression patterns. **Figure 3.4** show the expression profiles of 250 genes in 12 clusters between human and mouse. We therefore implemented time warping algorithm for each group of genes between human and mouse, and hypothesize that genes in the same cluster group have the same biological functions. **Table 3.2** clarifies the distance score and gene number of each cluster. At last, each cluster has its unique time points mapping pattern. In this approach, we want to find many different gene expression patterns on heart developmental stages and make a pilot study for the research of human and mouse heart development. Detailed information of each group after time-warping is provided in Appendix B. **Figure 3.5** clarifies the system flow of our approach using dynamic time-warping. **Figure 3.6** exhibits gene expression profiles and time-warping results in the 12 distinct clusters.

Table 3.2 12 clusters of genes and their numbers and scores in each distinct group.

Mouse Cluster	Human Cluster	Number of Genes	Score
M 1	H 1	11	30.2521
M 2	H 2	14	35.6842
M 3	H 3	16	34.3911
M 4	H 4	37	55.5629
M 5	H 5	3	11.5195
M 6	H 6	22	43.2182
M 7	H 7	35	50.6715
M 8	H 8	49	68.406
M 9	H 9	9	25.9278
M 10	H 10	38	53.0934
M 11	H 11	9	30.3447
M 12	H 12	7	26.919

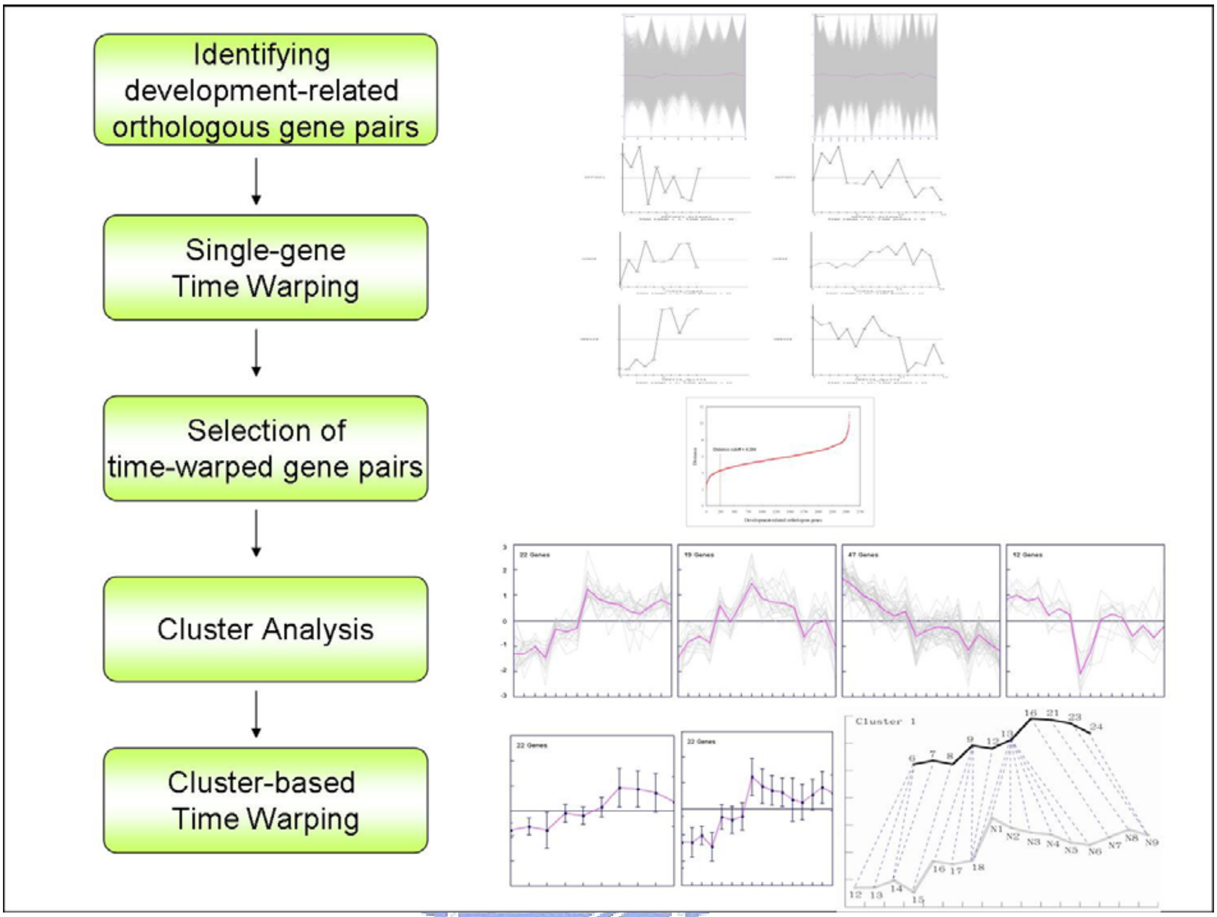
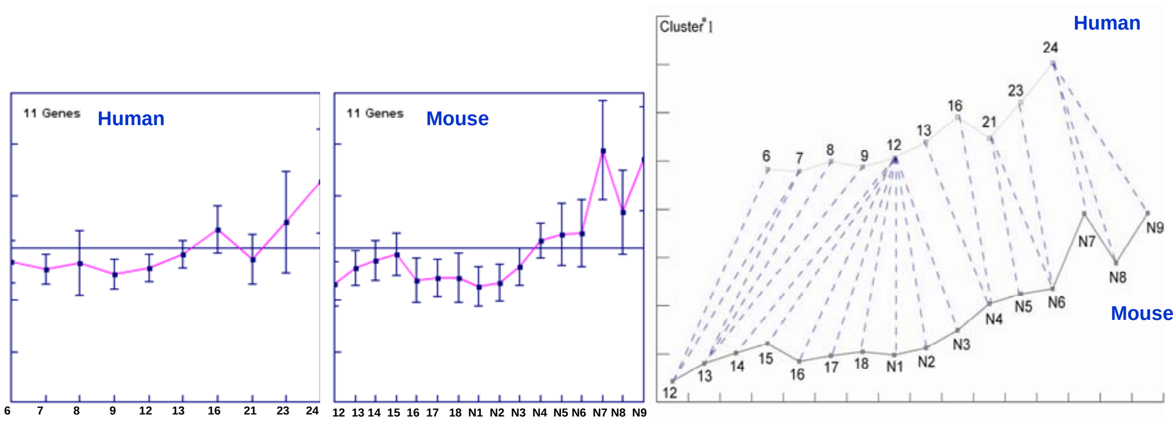
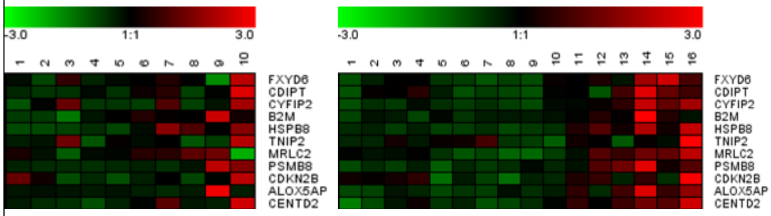

Identifying development-related orthologous gene pairs
Single-gene Time Warping
Selection of time-warped gene pairs
Cluster Analysis
Cluster-based Time Warping
22 Genes
19 Genes
47 Genes
12 Genes
22 Genes
22 Genes
Cluster 1

Figure 3.5 Flowchart of applying dynamic time-warping as a step.

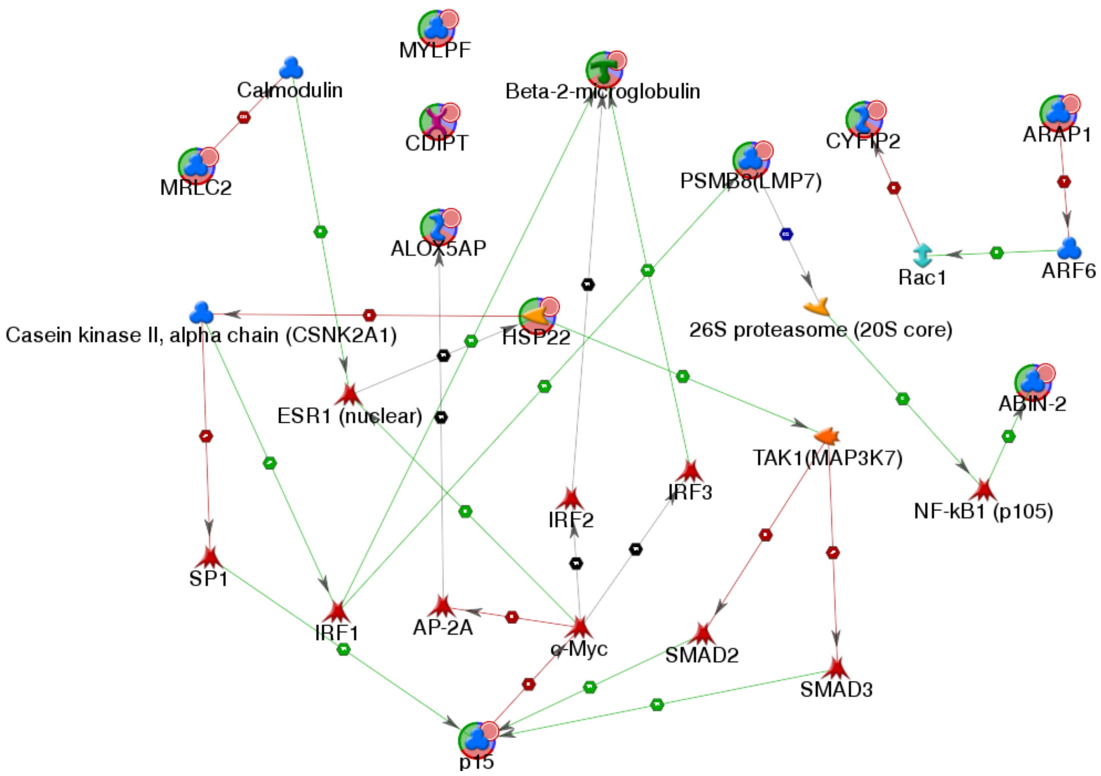
Firstly, single orthologous gene pair is applied the time warping step. After that, we selected the best genes which are warped great. Next step, the k-means clustering is used to clustering the best warped genes. The last step is to time-warp each cluster of genes individually.

Cluster 1

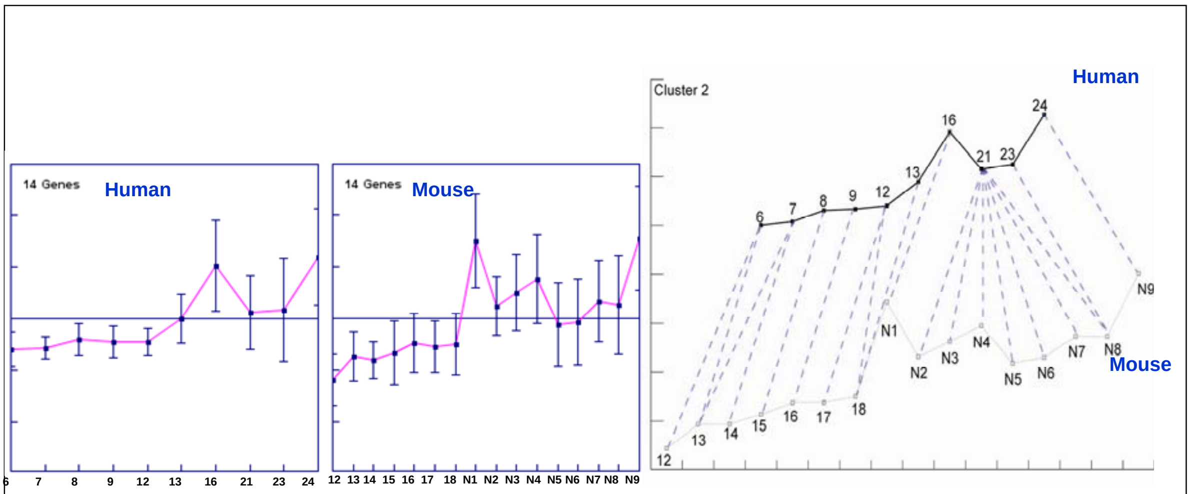


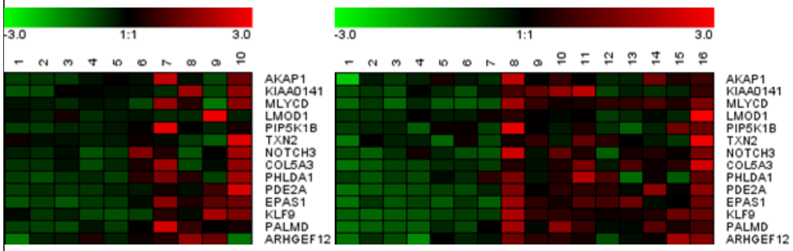


No. of human genes	11
No. of mouse genes	11
Distance	30.2521

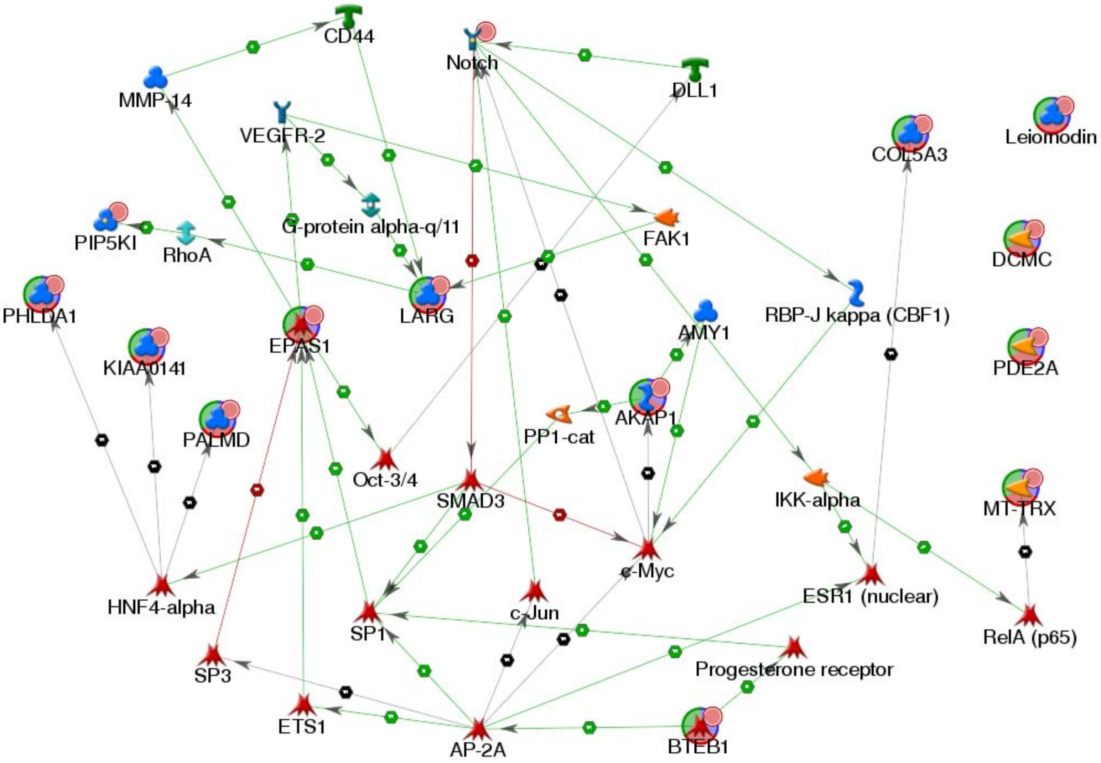


Cluster 2

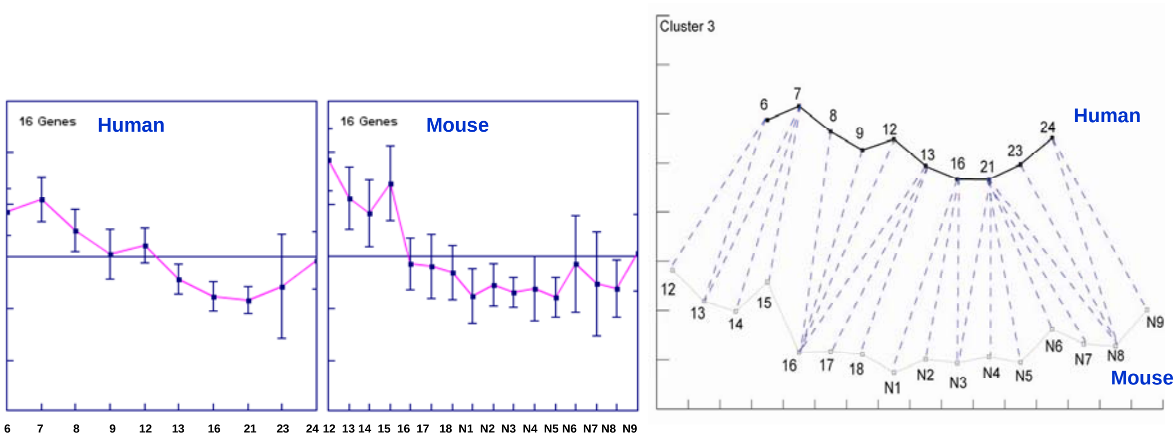


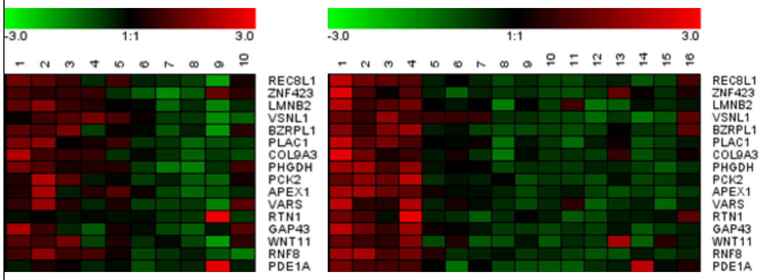


No. of human genes	14
No. of mouse genes	14
Distance	35.6842

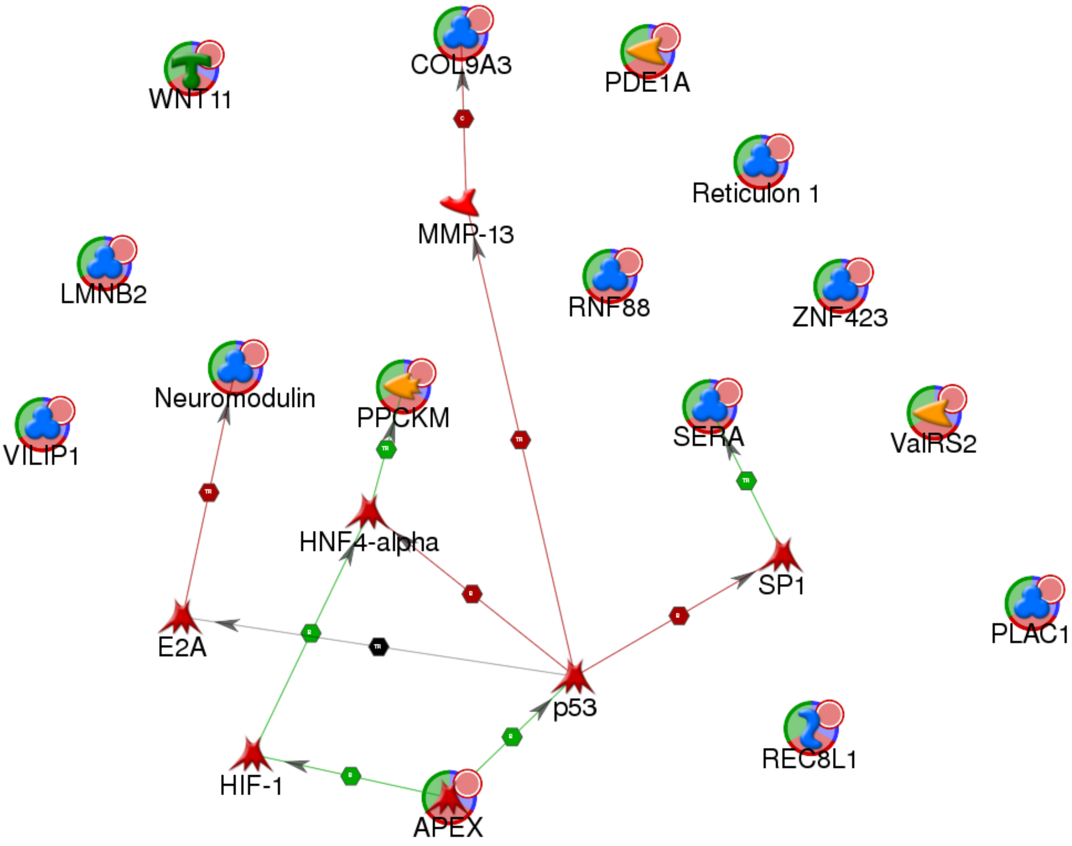


Cluster 3

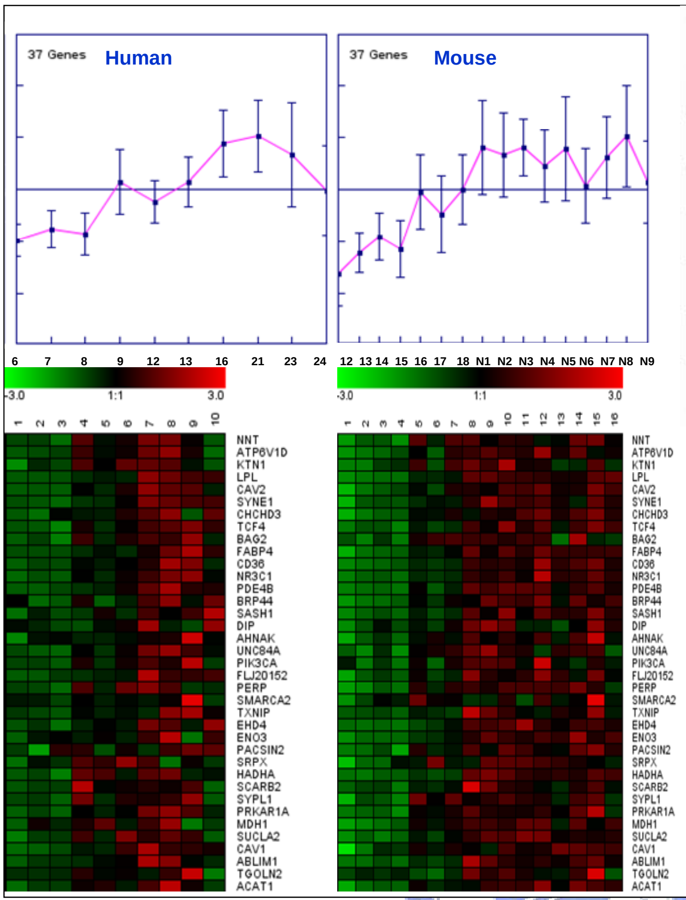


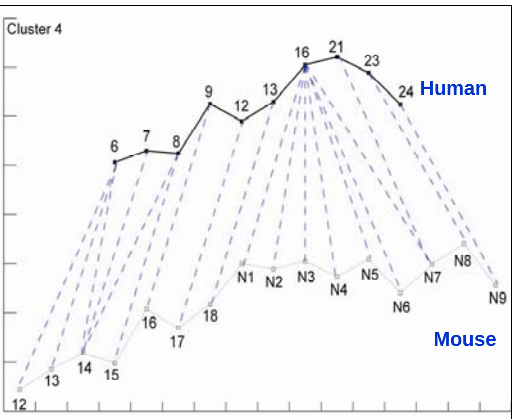


No. of human genes	16
No. of mouse genes	16
Distance	34.3911

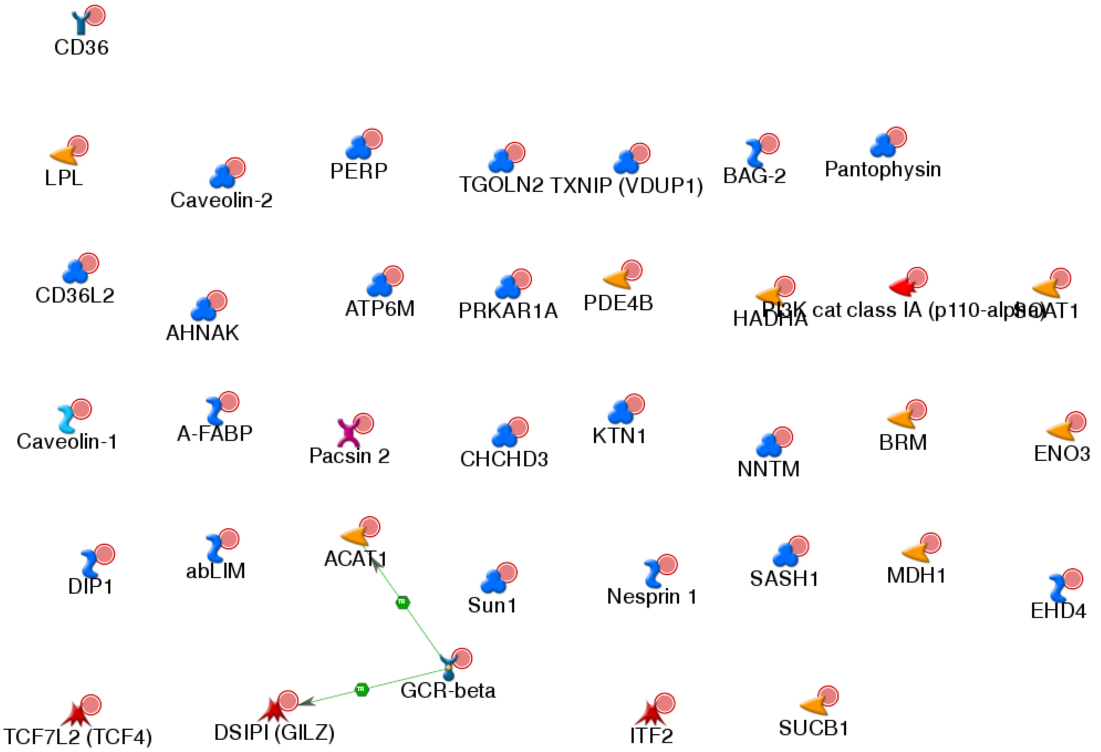


Cluster 4

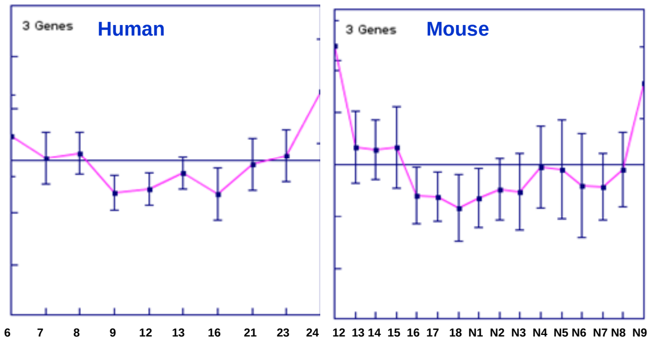


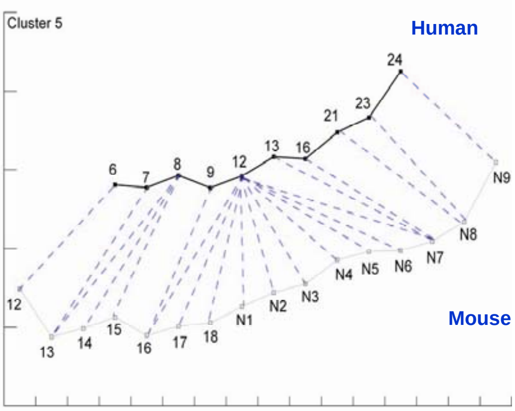


No. of human genes	37
No. of mouse genes	37
Distance	55.5629

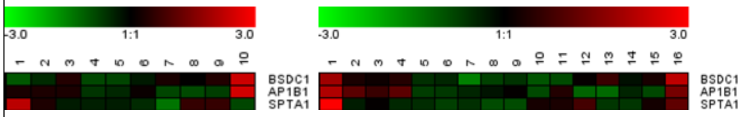


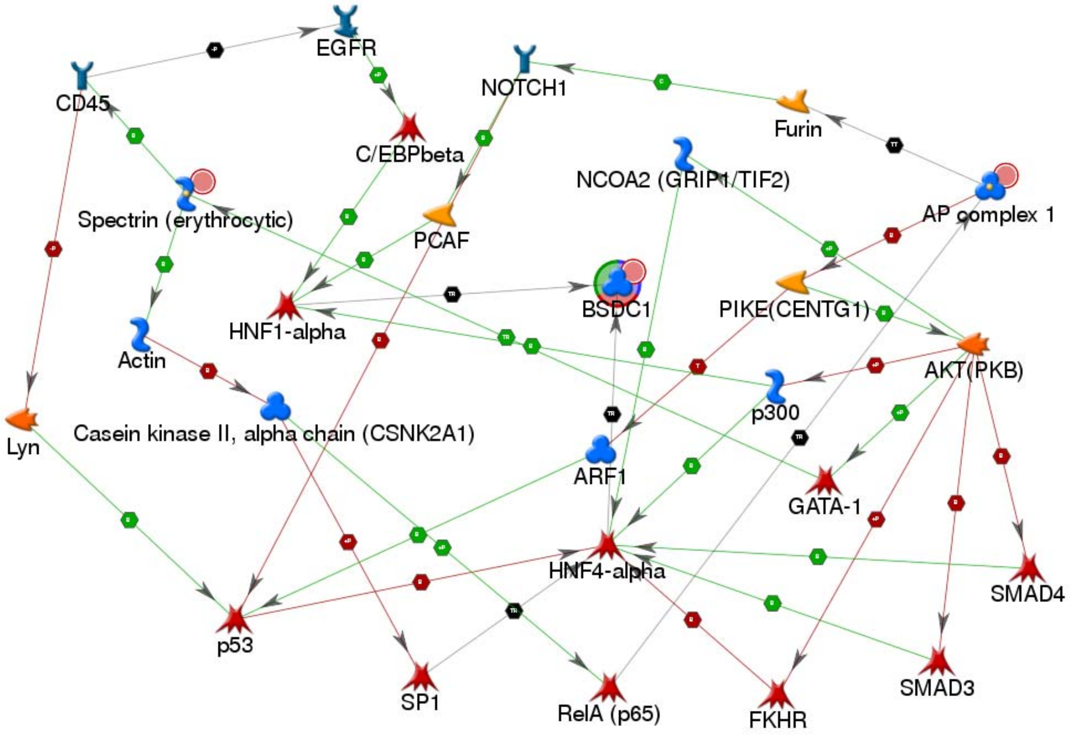
Cluster 5



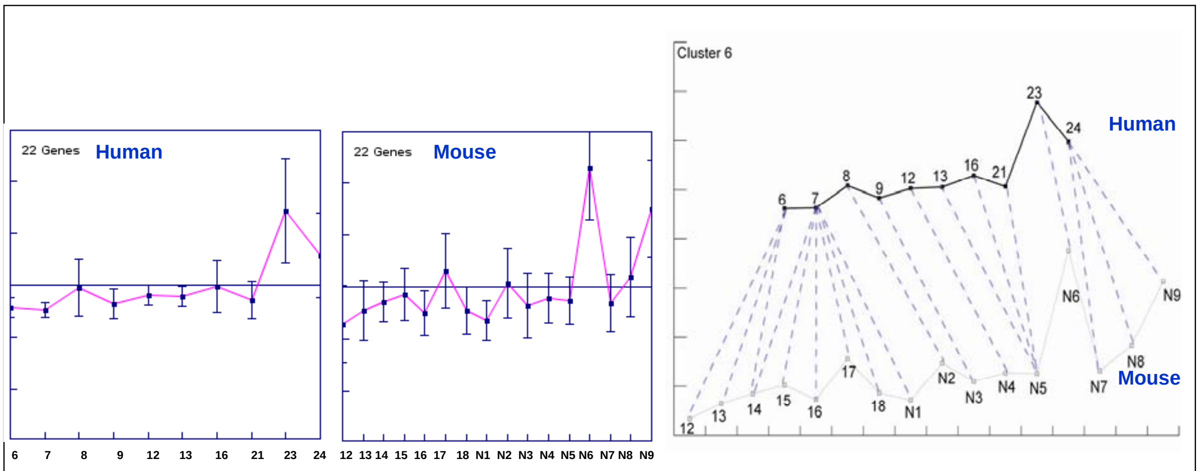


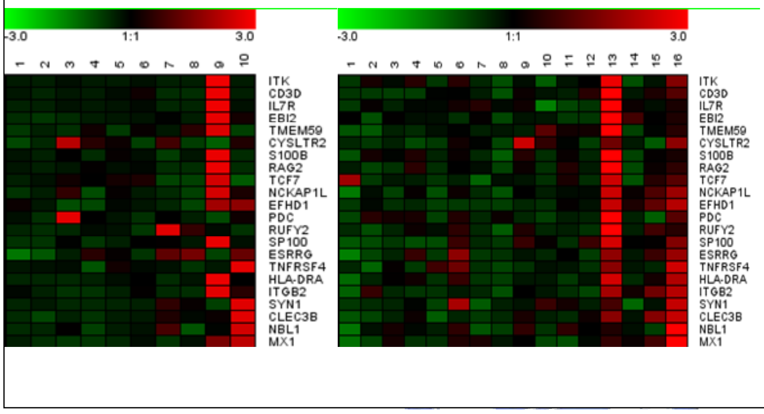
No. of human genes	3
No. of mouse genes	3
Distance	11.5195



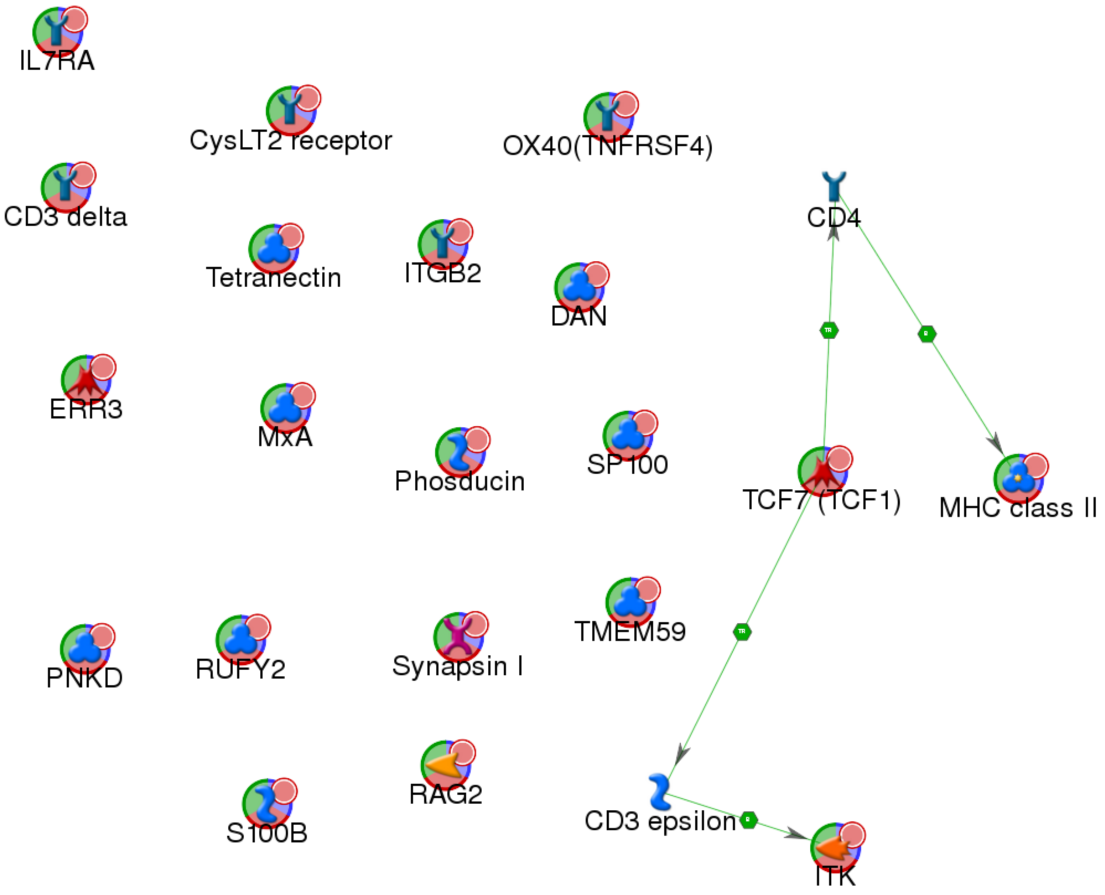


Cluster 6

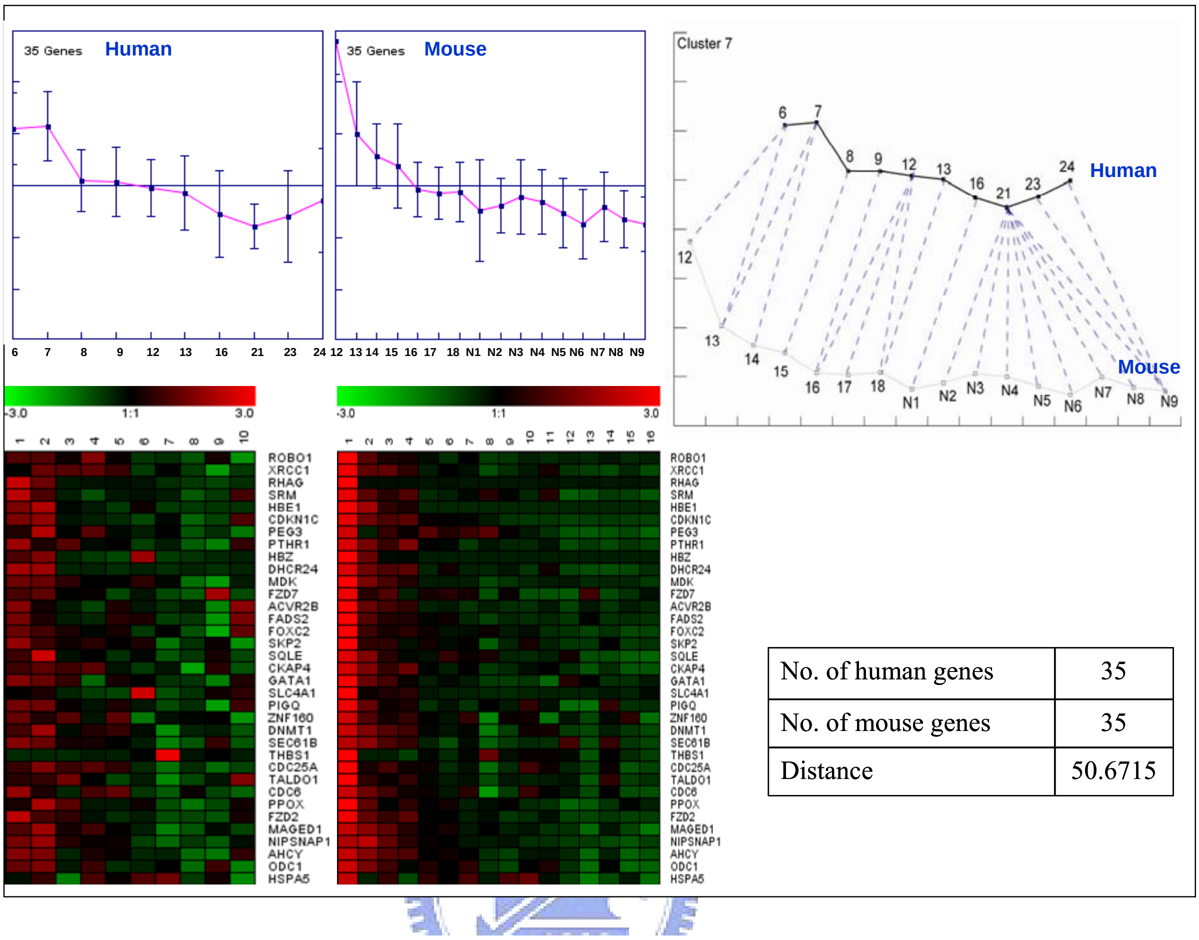




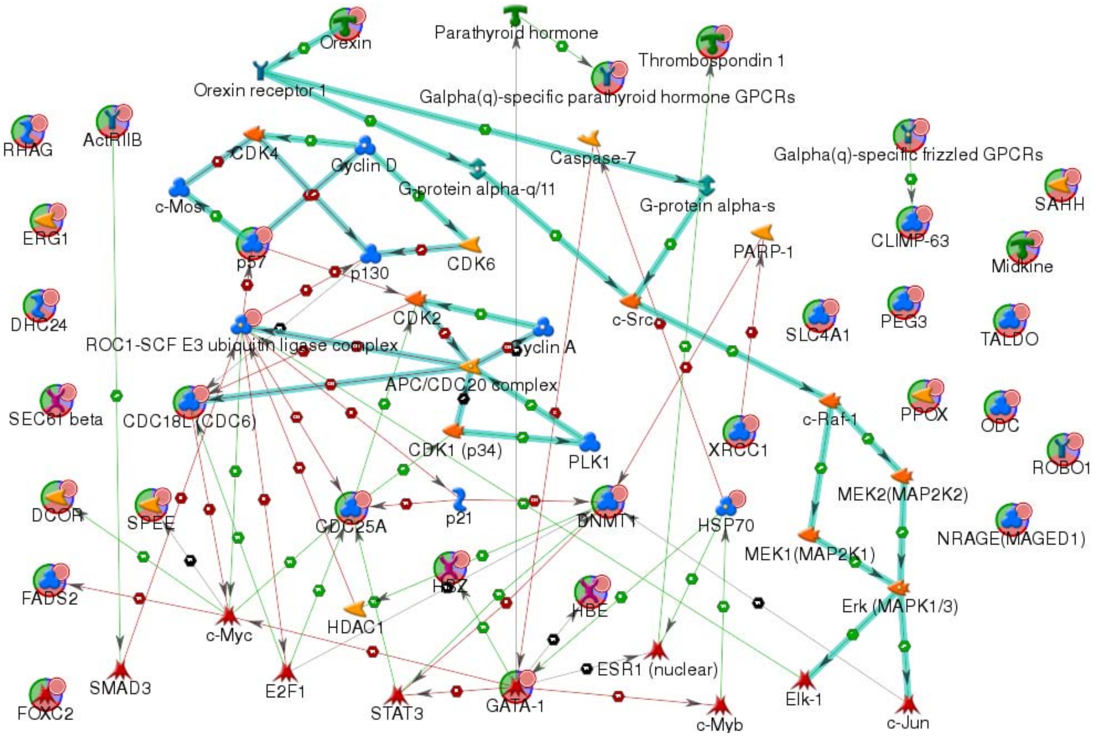
No. of human genes	22
No. of mouse genes	22
Distance	43.2182



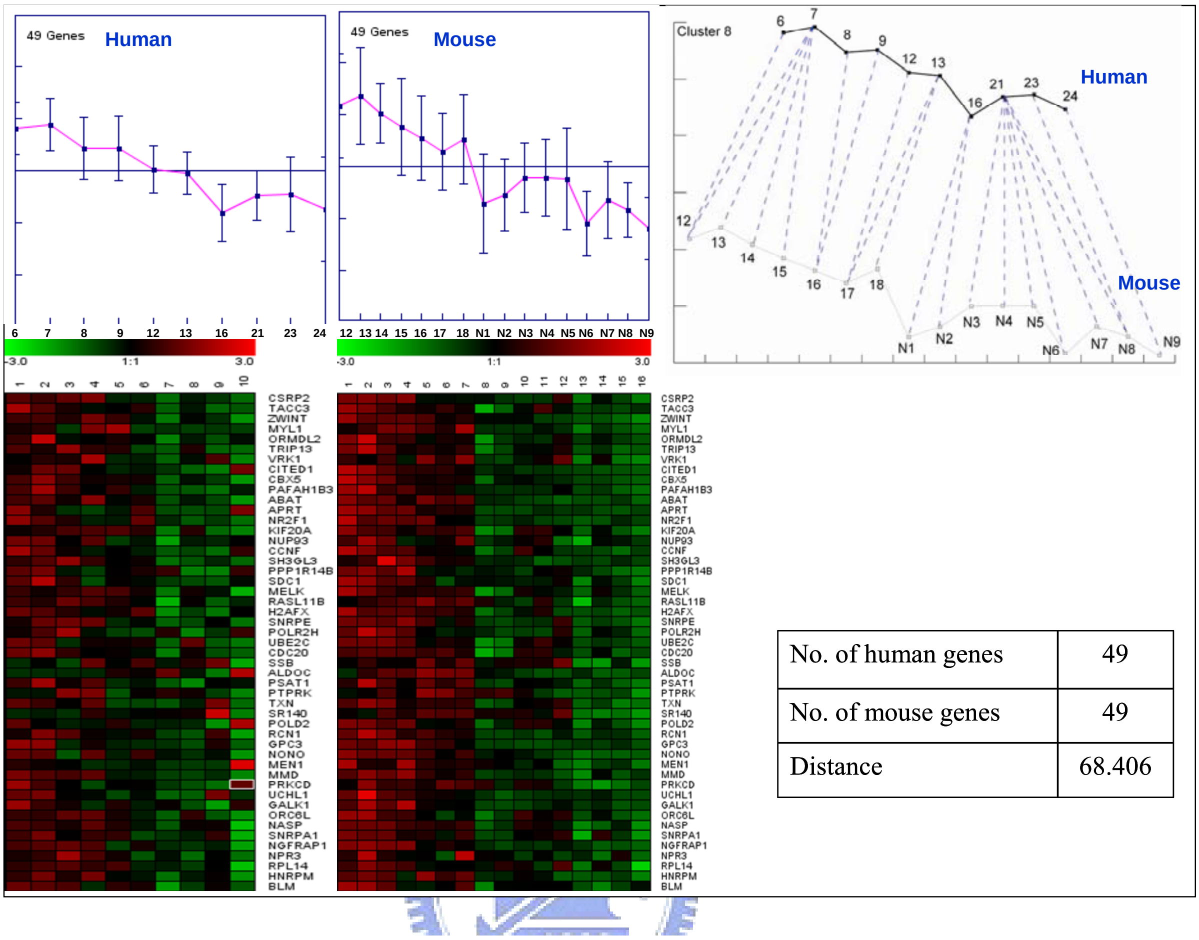
Cluster 7



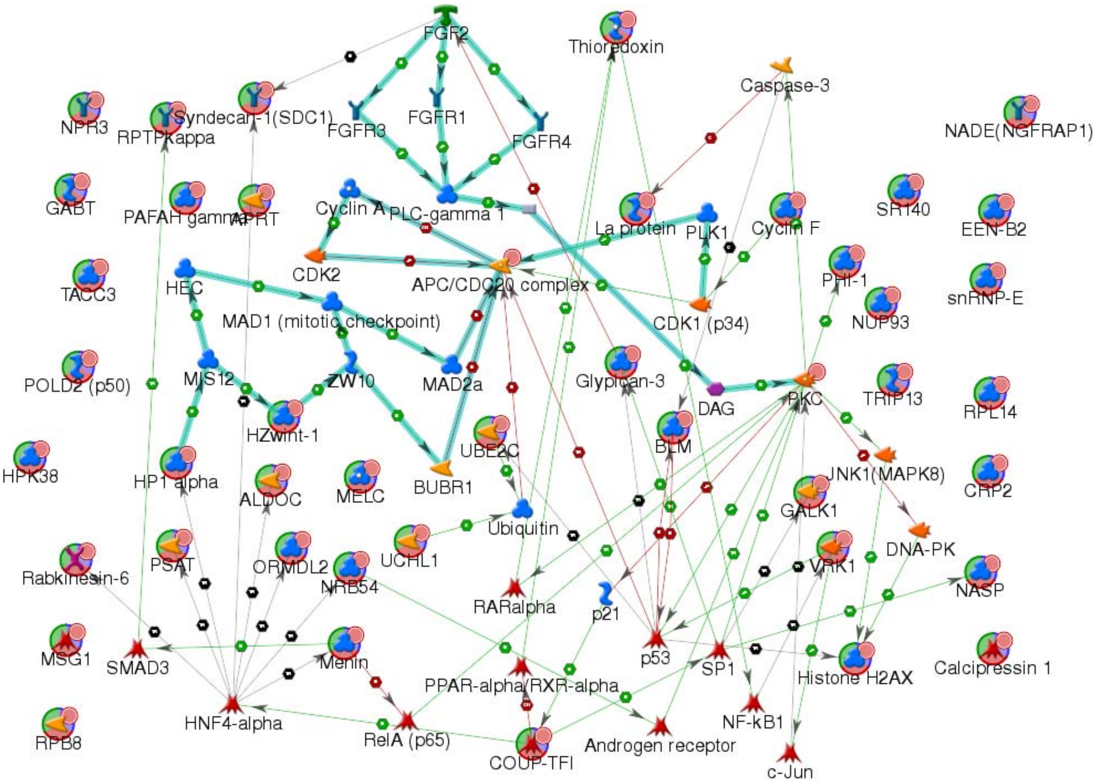
No. of human genes	35
No. of mouse genes	35
Distance	50.6715



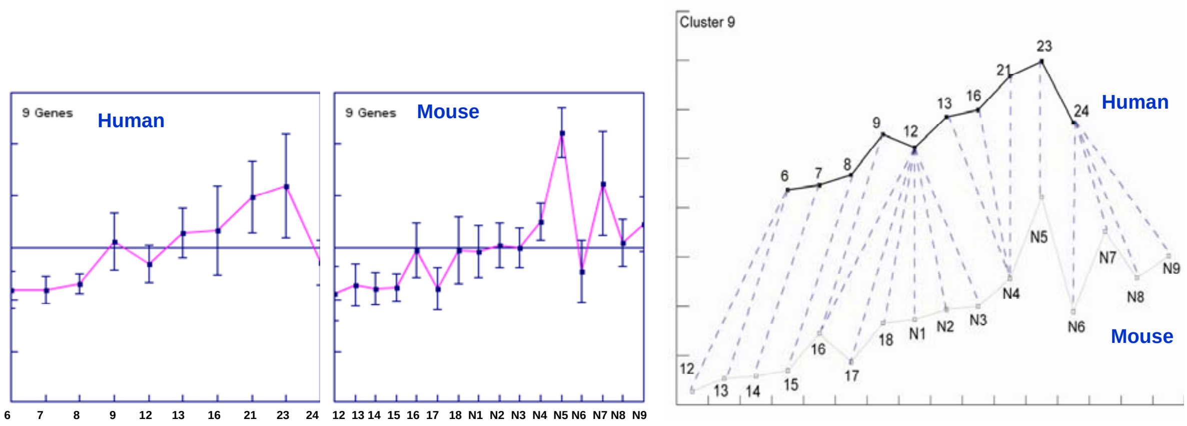
Cluster 8

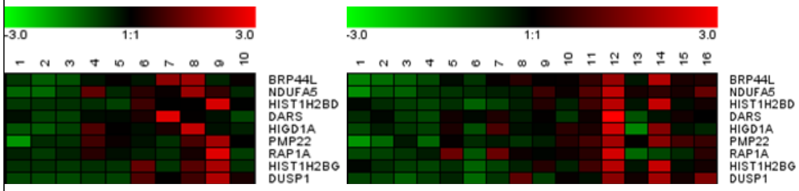


No. of human genes	49
No. of mouse genes	49
Distance	68.406

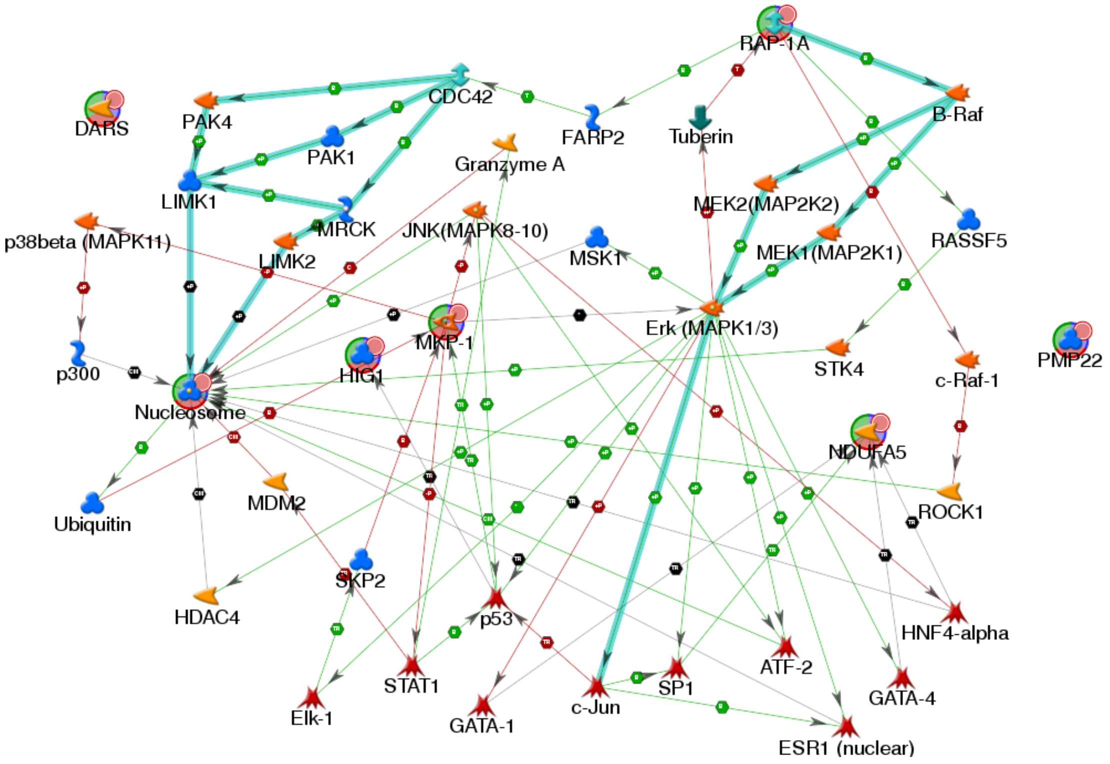


Cluster 9

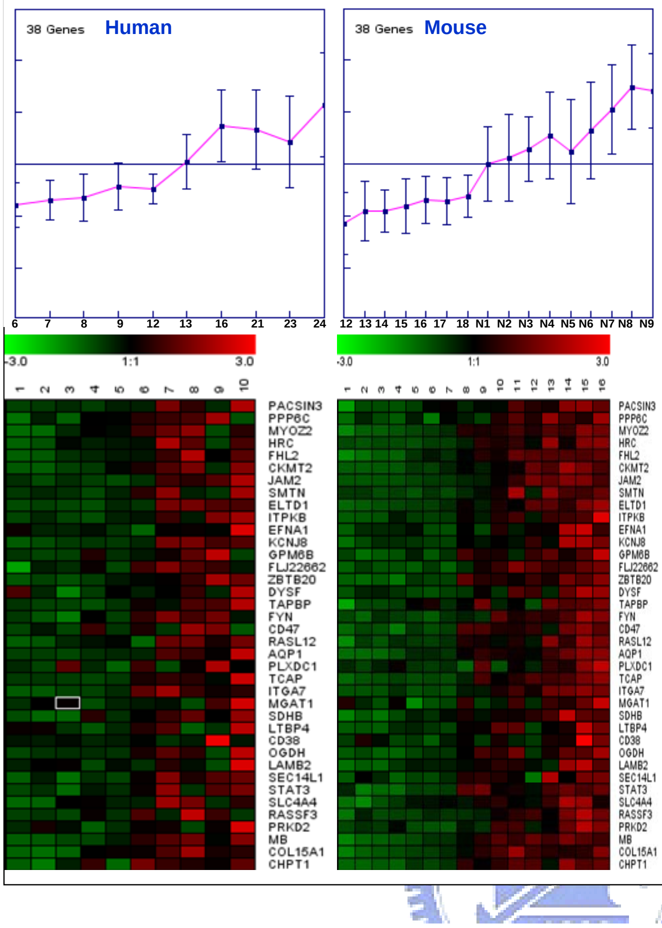


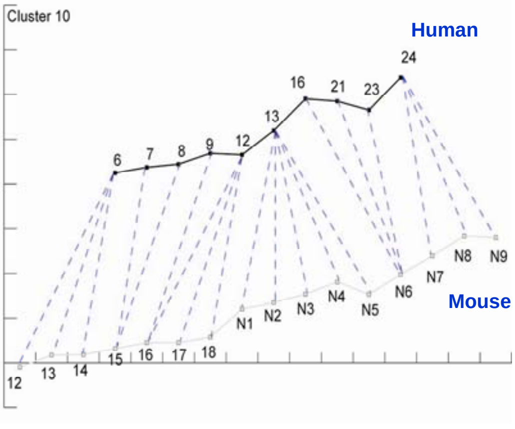


No. of human genes	9
No. of mouse genes	9
Distance	25.9278

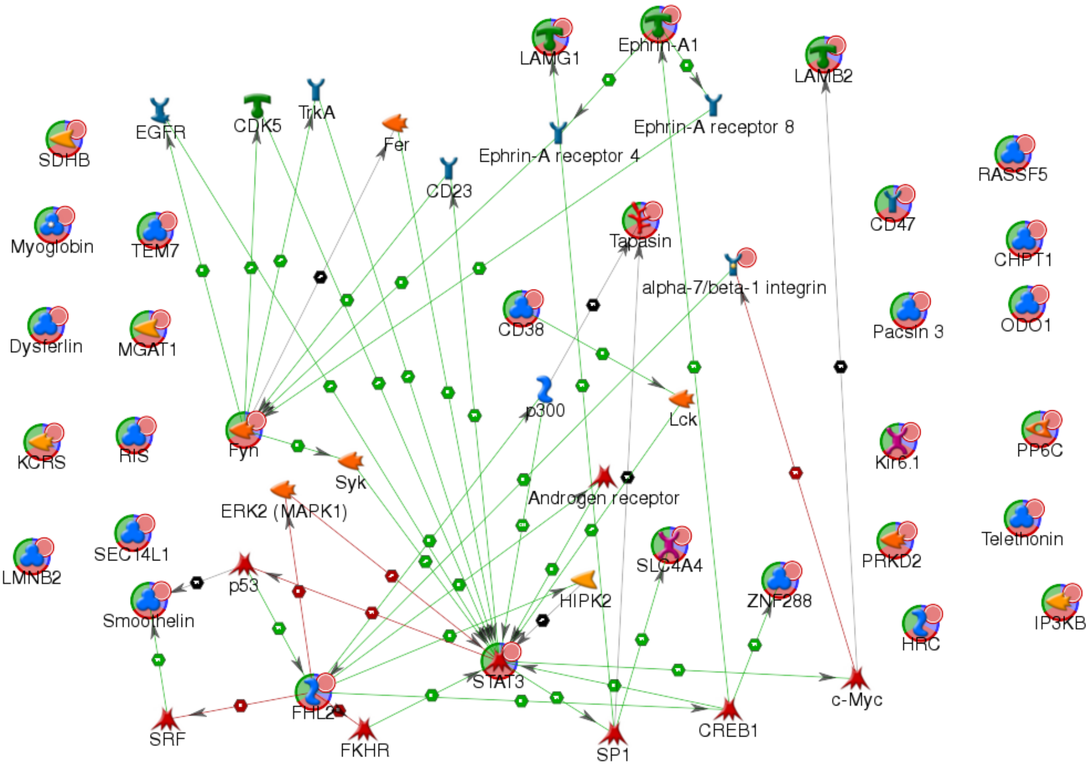


Cluster 10

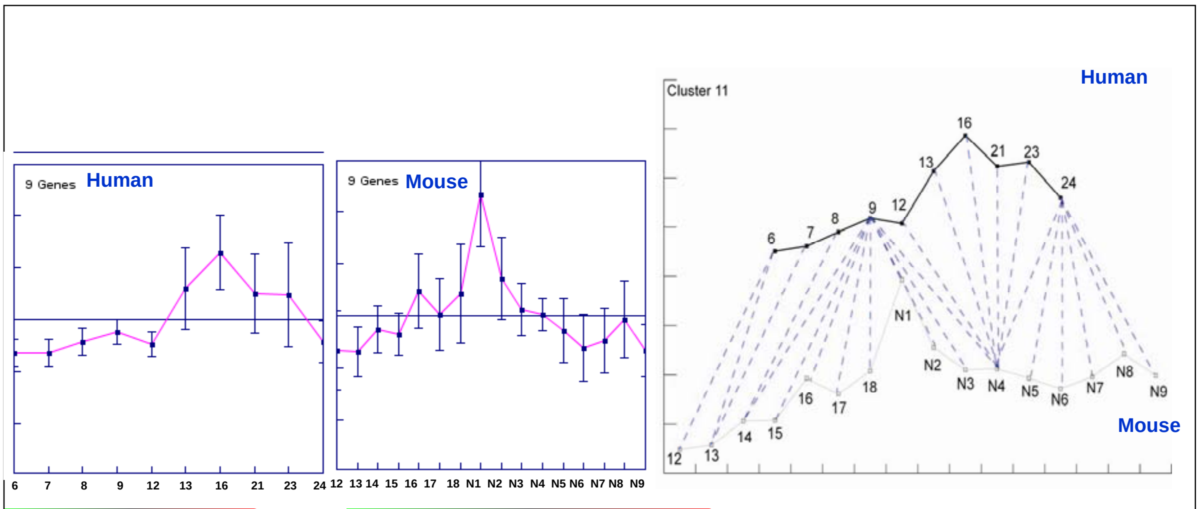


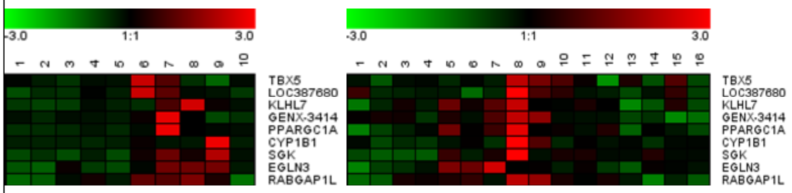


No. of human genes	38
No. of mouse genes	38
Distance	53.0934

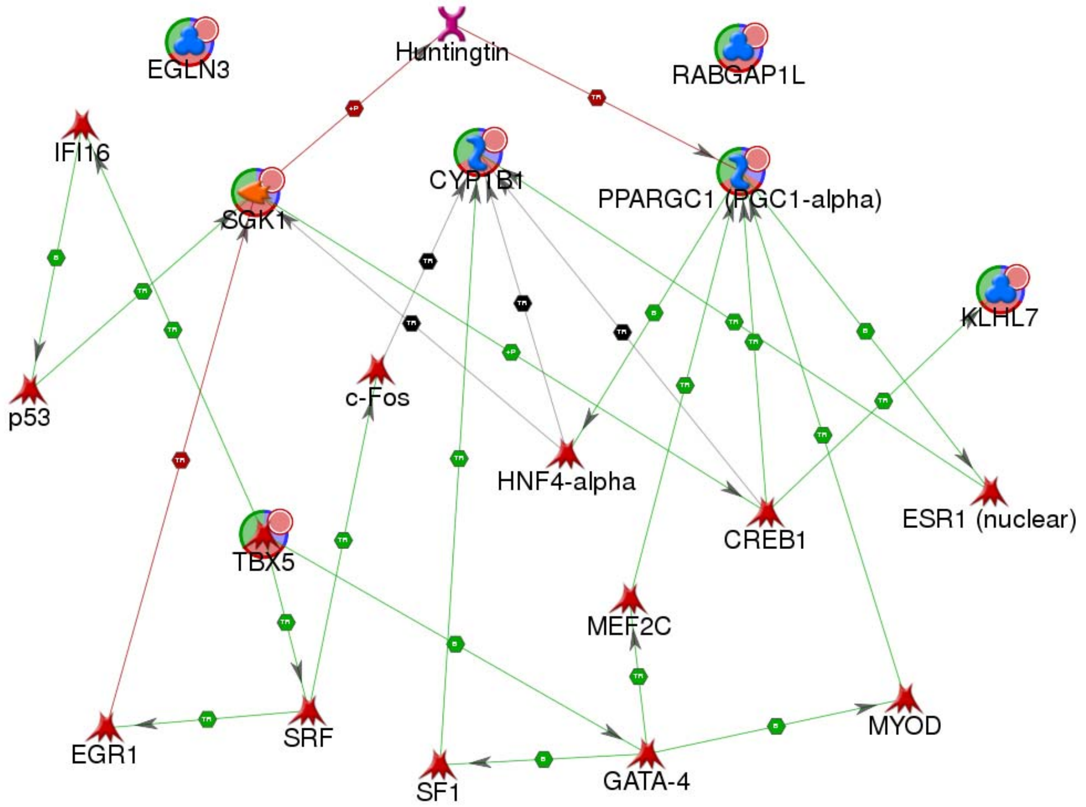


Cluster 11

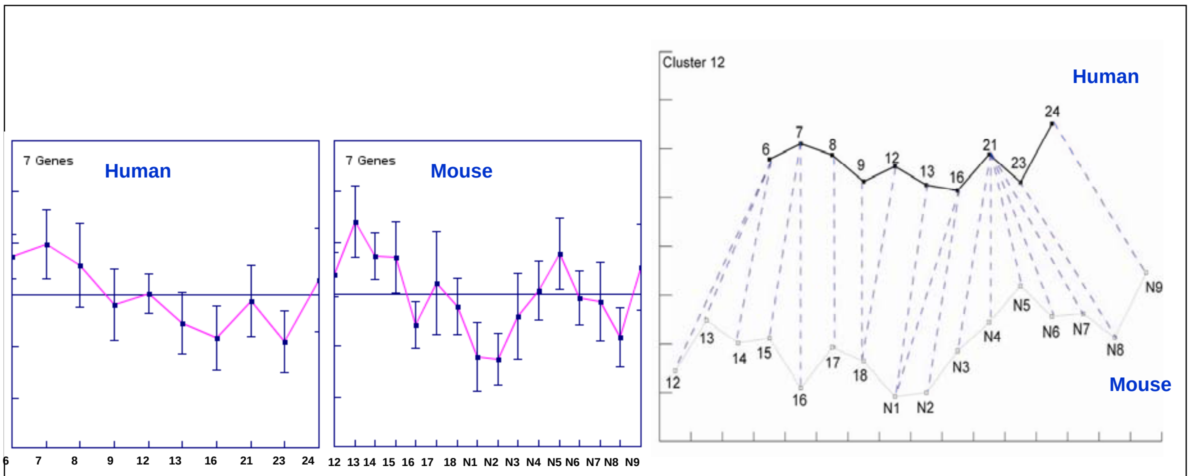


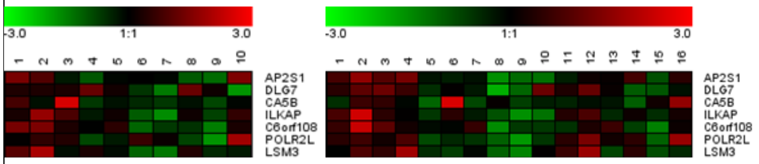


No. of human genes	9
No. of mouse genes	9
Distance	30.3447



Cluster 12





No. of human genes	7
No. of mouse genes	7
Distance	26.919

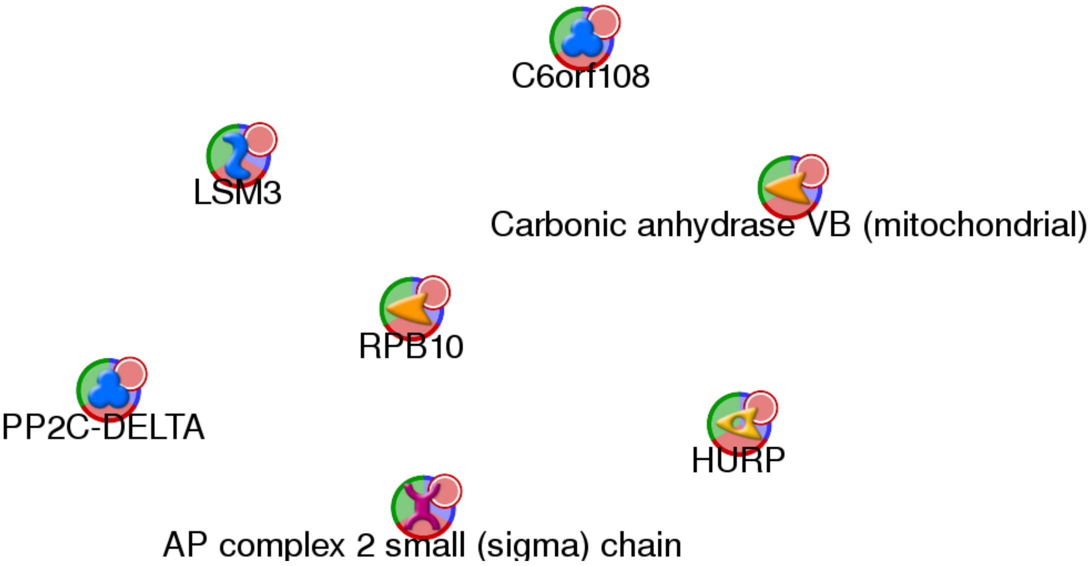

C6orf108
LSM3
Carbonic anhydrase VB (mitochondrial)
RPB10
PP2C-DELTA
HURP
AP complex 2 small (sigma) chain

Figure 3.6 Grouped time warping results and gene networks in the individual group.

After k-means clustering of the 250 genes, 12 clusters of genes, their expression and time-warping results are illustrated in each individual chart. Each chart displays the human and mouse expression profiles in the gene group on the top left. The expression value of each time point is the mean of all the genes in the cluster and standard deviation is also showed in the plot. Every value is estimated by the log ratio and each gene’s expression is also showed on the bottom left. The time-warping result of the cluster is on the right side. The gene network chart is below each time warping result chart. We used gene list in each individual cluster to make the network by applying MetaCore™ (a systematic software for analyzing microarray data) with shortest paths (Dijkstra’s shortest paths algorithm) to find the shortest directed paths between the grouped genes.

3.4 Functional Distribution of the Best 250 Time-Warped Genes

The gene-ontology database (GO: http://www.geneontology.org) is a useful tool for annotating and analyzing the function of large numbers of genes. Genes in GO are classified based on their annotated role in biological process, molecular functions, and cellular components. To determine which GO terms are more populated among the mapping genes, FatiGO[20]—a web-based application that facilitates GO terms querying—was used. Figure 6a shows GO biological process categories level-3 distribution of the best 250 time-warped genes. The most populated functional categories in humans and mice are cellular metabolic process, primary metabolic process, macromolecule metabolic process, regulation of biological process, cell communication, multicellular organismal development, anatomical structure development, and cellular developmental process. These populated categories are developmental-associated terms. Obviously, the 250 genes have large populations in the process of development.

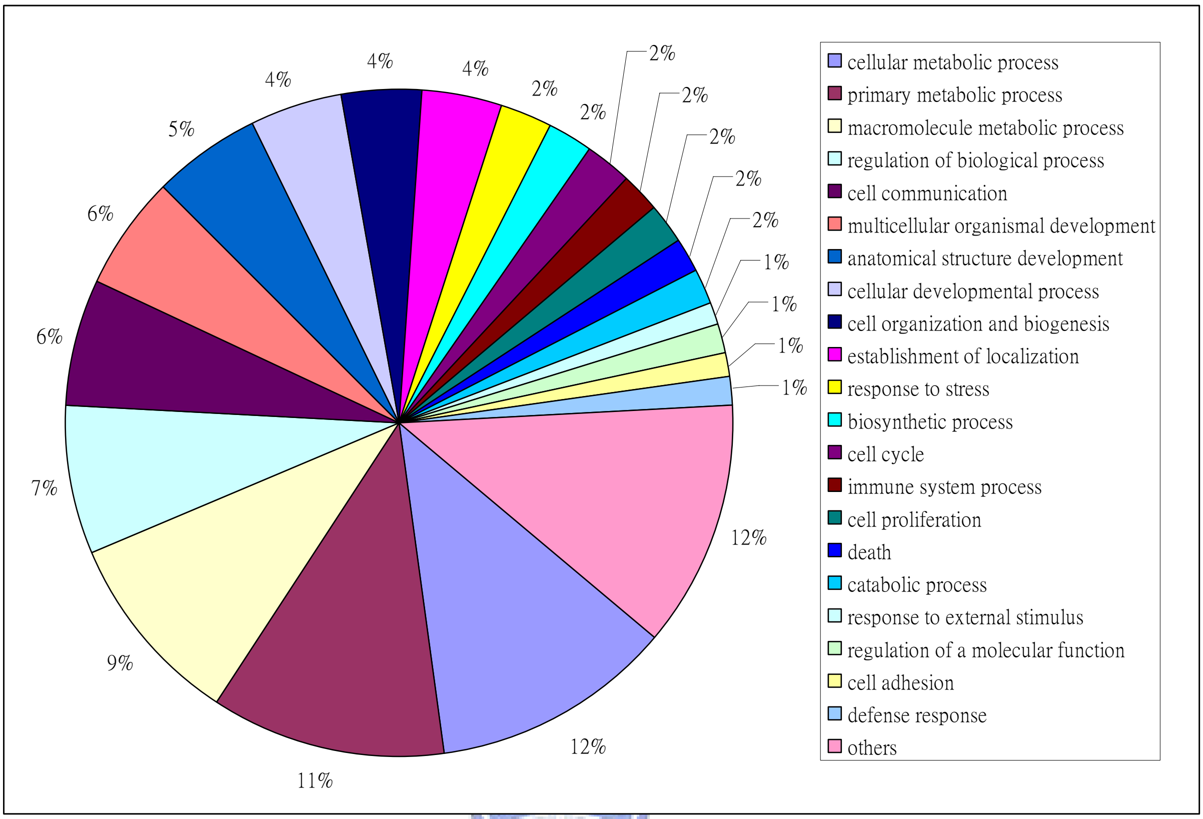


Figure 3.7 FatiGO result for the 250 genes in Level-3 Gene Ontology distribution.

The most populated GO categories are cellular metabolic process, primary metabolic process, macromolecule metabolic process, regulation of biological process, cell communication, multicellular organismal development, anatomical structure development, and cellular developmental process, etc. The distribution of the mapping functional categories is clearly shown, as some important categories associated with development.

3.5 Finding Statistically Overrepresented GO terms

To investigate the biological functions involved in human and mouse time-warping genes, the GO categories were analyzed using the GeneGO web-based program. GeneGO calculates statistical significance of nonrandom representations, that is, enrichment of a GO category among the gene under investigation. The nonrandom enrichment of a variety of biological process categories were identified, including organ development, cell differentiation, cellular developmental process, system development, developmental process, cell development, etc. These GO categories were statistically significant ($p < 0.005$) with genes in the microarray chip for humans and mice. Interestingly, the significant GO terms are highly correlated with embryo development. Unequivocally, this overrepresented GO analysis validates the orthologous time-warping system and the microarray gene expression profiles are useful for studying vertebrate embryonic development. Additionally, selected time-warped genes also demonstrated enriched annotations related to cellular components, including extracellular matrix and molecular functions such as hydrolase activity and growth factor activity. These biological gene categories enriched in 250 genes can provide direction for future investigations into the molecular mechanisms of heart development. **Table 3.3** presents the significant GO terms in total 250 time-warped genes and individual cluster. We selected 12 GO categories that are the most significant in each dataset. As shown in **Table 3.3**, genes in cluster 4 are overrepresented most in transcription and metabolic process. Genes in cluster 6 are overrepresented most in immune system process, lymphocyte differentiation, T cell differentiation. Genes in cluster 7 are overrepresented most in cell cycle process. Genes in cluster 10 are overrepresented most in signaling pathway and system development.

3.5.1 P-value Function

The P-value Function:

$$z\text{-}score = \frac{r - n\frac{R}{N}}{\sqrt{n\left(\frac{R}{N}\right)\left(1 - \frac{R}{N}\right)\left(1 - \frac{n-1}{N-1}\right)}}$$

The P-value is calculated using the same basic formula: a hypergeometric distribution where the P-value essentially represents the probability of particular mapping arising by chance, given the numbers of genes in the set of all genes on processes, genes on a particular process and genes in datasets. This function uses the same variables as the Z-Score.

Variables:

N - total number of nodes in MetaCore database

R - number of the network's objects corresponding to the genes and proteins in user’s list

n - total number of nodes in each small network generated from user’s list

r - number of nodes with data in each small network generated from user’s list

Table 3.3 Biological process that Gene Ontology categories non-randomly enrich in 250 time-warping genes and individual clusters.

Cluster	Process	Percentage	P-values
250 genes (Cluster1-Cluster12)	positive regulation of biological process	36.34	1.79E-25
	biological regulation	63.87	2.77E-25
	regulation of cellular process	55.04	3.03E-25

	regulation of biological process	61.34	3.11E-25
	organ development	36.34	9.58E-24
	positive regulation of cellular process	31.3	8.50E-23
	cell differentiation	43.28	1.20E-22
	cellular developmental process	43.28	1.20E-22
	signal transduction	49.58	6.84E-22
	system development	41.18	1.49E-20
	developmental process	58.19	2.33E-20
	cell development	36.76	1.10E-19
Cluster1	regulation of Rho protein signal transduction	13.64	5.15E-06
	negative regulation of receptor mediated endocytosis	9.09	6.21E-06
	positive regulation of metabolic process	40.91	6.94E-06
	transcription from RNA polymerase II promoter	45.45	1.09E-05
	paraxial mesoderm morphogenesis	9.09	1.86E-05
	regulation of Ras protein signal transduction	13.64	2.82E-05
	positive regulation of transcription, DNA-dependent	31.82	3.13E-05
	paraxial mesoderm development	9.09	3.72E-05
	ruffle organization and biogenesis	9.09	3.72E-05
	positive regulation of transcription from RNA polymerase II promoter	27.27	5.55E-05
	regulation of small GTPase mediated signal transduction	13.64	5.59E-05
	regulation of transcription from RNA polymerase II promoter	36.36	7.41E-05
Cluster2	anatomical structure morphogenesis	59.38	2.51E-09
	anatomical structure development	75	1.01E-08
	regulation of transcription, DNA-dependent	53.12	1.12E-08
	regulation of transcription	56.25	1.57E-08
	organ development	62.5	2.03E-08
	positive regulation of transcription	37.5	3.13E-08
	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	56.25	4.41E-08
	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	37.5	4.51E-08
	positive regulation of transcription, DNA-dependent	34.38	6.55E-08
	regulation of cellular metabolic process	59.38	7.69E-08
	transcription, DNA-dependent	53.12	1.29E-07
	RNA biosynthetic process	53.12	1.35E-07

Cluster3	base-excision repair	12.5	8.70E-04
	positive regulation of transcription from RNA polymerase II promoter	25	1.49E-03
	regulation of helicase activity	6.25	1.86E-03
	negative regulation of helicase activity	6.25	1.86E-03
	protein import into nucleus, translocation	12.5	2.54E-03
	intracellular protein transport across a membrane	12.5	2.54E-03
	regulation of mitochondrial membrane permeability	6.25	3.71E-03
	positive regulation of transcription, DNA-dependent	25	4.60E-03
	response to hypoxia	12.5	7.04E-03
	positive regulation of global transcription from RNA polymerase II promoter	6.25	7.40E-03
	response to X-ray	6.25	7.40E-03
	response to stress	37.5	7.71E-03
Cluster4	regulation of transcription	50	6.61E-11
	regulation of cellular metabolic process	55.36	6.97E-11
	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	50	3.09E-10
	regulation of metabolic process	55.36	3.81E-10
	transcription	50	1.96E-09
	regulation of transcription, DNA-dependent	41.07	1.71E-08
	transcription from RNA polymerase II promoter	35.71	6.22E-08
	transcription, DNA-dependent	42.86	7.36E-08
	RNA biosynthetic process	42.86	7.86E-08
	positive regulation of transcription	26.79	1.06E-07
	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	26.79	1.63E-07
	positive regulation of cellular metabolic process	28.57	2.94E-07
Cluster5	positive regulation of cellular metabolic process	48.15	2.74E-09
	positive regulation of metabolic process	48.15	5.29E-09
	positive regulation of transcription	40.74	4.31E-08
	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	40.74	6.04E-08
	positive regulation of cellular process	59.26	7.35E-08
	positive regulation of transcription from RNA polymerase II promoter	33.33	1.05E-07
	positive regulation of transcription,	37.04	1.17E-07

	DNA-dependent		
	positive regulation of biological process	59.26	9.61E-07
	regulation of cellular metabolic process	55.56	5.88E-06
	regulation of transcription from RNA polymerase II promoter	37.04	7.57E-06
	regulation of cell proliferation	37.04	8.27E-06
	regulation of cellular process	74.07	1.20E-05
Cluster6	immune response	63.16	2.35E-10
	immune system process	68.42	1.09E-09
	T cell differentiation	26.32	1.90E-07
	T cell activation	31.58	2.53E-07
	cell activation	36.84	6.41E-07
	lymphocyte differentiation	26.32	2.04E-06
	lymphocyte activation	31.58	2.96E-06
	T cell selection	15.79	4.99E-06
	response to stimulus	73.68	6.49E-06
	leukocyte activation	31.58	7.01E-06
	leukocyte differentiation	26.32	1.42E-05
	multicellular organismal process	89.47	1.52E-05
Cluster7	regulation of progression through cell cycle	40.91	9.43E-18
	regulation of cell cycle	40.91	1.18E-17
	cell cycle process	43.94	3.50E-16
	cell cycle	43.94	1.38E-15
	mitotic cell cycle	30.3	2.33E-14
	G1 phase of mitotic cell cycle	13.64	3.84E-14
	G1 phase	13.64	6.10E-14
	interphase	22.73	7.97E-14
	interphase of mitotic cell cycle	22.73	7.97E-14
	cell cycle phase	30.3	1.66E-13
	biological regulation	81.82	9.82E-12
	regulation of biological process	75.76	7.22E-10
Cluster8	regulation of progression through cell cycle	30.99	8.12E-12
	regulation of cell cycle	30.99	9.62E-12
	regulation of mitosis	14.08	3.38E-11
	cell cycle	35.21	4.76E-11
	cell cycle process	32.39	6.90E-10
	mitosis	16.9	1.28E-09
	M phase of mitotic cell cycle	16.9	1.39E-09
	mitotic checkpoint	8.45	1.24E-08

	cyclin catabolic process	5.63	2.10E-08
	M phase	16.9	1.21E-07
	regulation of exit from mitosis	7.04	1.33E-07
	mitotic sister chromatid segregation	8.45	1.57E-07
Cluster9	intracellular signaling cascade	52.08	4.98E-11
	protein amino acid phosphorylation	35.42	1.11E-10
	cell differentiation	66.67	2.58E-10
	cellular developmental process	66.67	2.58E-10
	cell development	60.42	5.03E-10
	nucleosome assembly	14.58	7.75E-10
	biopolymer metabolic process	75	8.20E-10
	phosphorylation	35.42	2.04E-09
	chromatin assembly	14.58	7.55E-09
	phosphate metabolic process	35.42	3.07E-08
	phosphorus metabolic process	35.42	3.07E-08
	developmental process	77.08	3.41E-08
Cluster10	positive regulation of biological process	42.22	6.34E-05
	positive regulation of cellular process	37.78	7.87E-05
	integrin-mediated signaling pathway	8.89	1.25E-04
	signal transduction	55.56	1.97E-04
	protein amino acid autophosphorylation	8.89	2.39E-04
	cell communication	60	2.53E-04
	regulation of biological quality	24.44	2.65E-04
	protein autoprocessing	8.89	2.89E-04
	system development	46.67	3.29E-04
	organ development	40	3.47E-04
	protein amino acid phosphorylation	20	4.36E-04
	immune response-activating cell surface receptor signaling pathway	6.67	4.41E-04
Cluster11	transcription from RNA polymerase II promoter	77.78	7.47E-12
	regulation of transcription, DNA-dependent	77.78	1.43E-10
	regulation of transcription from RNA polymerase II promoter	66.67	1.60E-10
	positive regulation of transcription, DNA-dependent	55.56	8.67E-10
	regulation of transcription	77.78	1.23E-09
	transcription, DNA-dependent	77.78	1.27E-09
	RNA biosynthetic process	77.78	1.33E-09
	positive regulation of transcription from RNA polymerase II promoter	50	1.41E-09

	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	77.78	2.92E-09
	positive regulation of transcription	55.56	4.15E-09
	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	55.56	5.71E-09
	transcription	77.78	8.36E-09
Cluster12	mitotic chromosome movement towards spindle pole	20	1.16E-03
	positive regulation of mitotic metaphase/anaphase transition	20	1.16E-03
	chromosome movement towards spindle pole	20	1.16E-03
	clathrin cage assembly	20	1.74E-03
	positive regulation of mitosis	20	2.90E-03
	membrane budding	20	3.47E-03
	vesicle coating	20	3.47E-03
	regulation of transcription from RNA polymerase I promoter	20	4.05E-03
	regulation of mitotic metaphase/anaphase transition	20	4.63E-03
	establishment of chromosome localization	20	5.21E-03
	chromosome localization	20	5.21E-03
	mitotic metaphase/anaphase transition	20	5.79E-03

3.6 Analysis of Transcriptional Regulations

3.6.1 Transcription Factors in Clusters

There are 14 genes act as transcription factors in selected 250 time-warped genes, the detailed information is listed in **Table 3.4**. It is obvious that these transcription factors are co-expressed with their corresponding cluster genes. For example, APEX1 is a TF in cluster 3. That means APEX1 are highly correlated with 15 genes in cluster 3. Genes which expressions are similar clustered into the same cluster. If there has any TF in each cluster, we may hypothesize that maybe the TF regulates the genes in the same cluster.

Table 3.4 Transcription factors in each cluster.

Cluster	Human	Mouse	Genes	TF
2	KLF9	Klf9	14	2
	EPAS1	Epas1		
3	APEX1	Apex1	16	1
4	DIP	BC021523	37	2
	TCF4	Tcf4		

6	ESRRG	Esrrg	22	2
	TCF7	Tcf7		
7	FOXC2	Foxc2	35	2
	GATA1	Gata1		
8	NR2F1	Nr2f1	49	3
	RCN1	Rcn1		
	CITED1	Cited1		
10	STAT3	Stat3	38	1
11	TBX5	Tbx5	9	1

3.6.2 Transcription Factors Regulations

In cluster 4, DIP and TCF4 are two transcription factors in total 37 genes. Their expression was shown in **Figure 3.8**. The expression profiles of these two genes are very similar in mouse heart development, but in human, DIP is dramatically degraded in latter time points and up-regulated in the latest time point. This condition is contrary to TCF4. These two TFs have similar pattern after time-warping between human and mouse. It is suggested that DIP and TCF4 maybe regulate the genes of the cluster. We can see the same condition in **Figure 3.9** (cluster 7), **Figure 3.10** (cluster 8), **Figure 3.11** (cluster 10). In mouse cluster 7, the two TFs, Foxc2 and Gata1, are dramatically down-regulated in the first two time points and smoothly expressed in latter points. In human cluster 7, FOXC2 and GATA1 mostly degraded in latter points. In the case, there is an interesting finding that the development rate or biological mechanism is different between human and mouse. In cluster 7 and cluster 8, the gene expression degraded through the time series, but in their networks, the regulation mechanisms are different. From the result, these two clusters have different regulators control their expressions. In cluster 10, STAT3 in an important factor in signaling transduction and regulated many genes in this cluster. We may suggest that STAT3 regulated other genes in this cluster not yet validated in public.

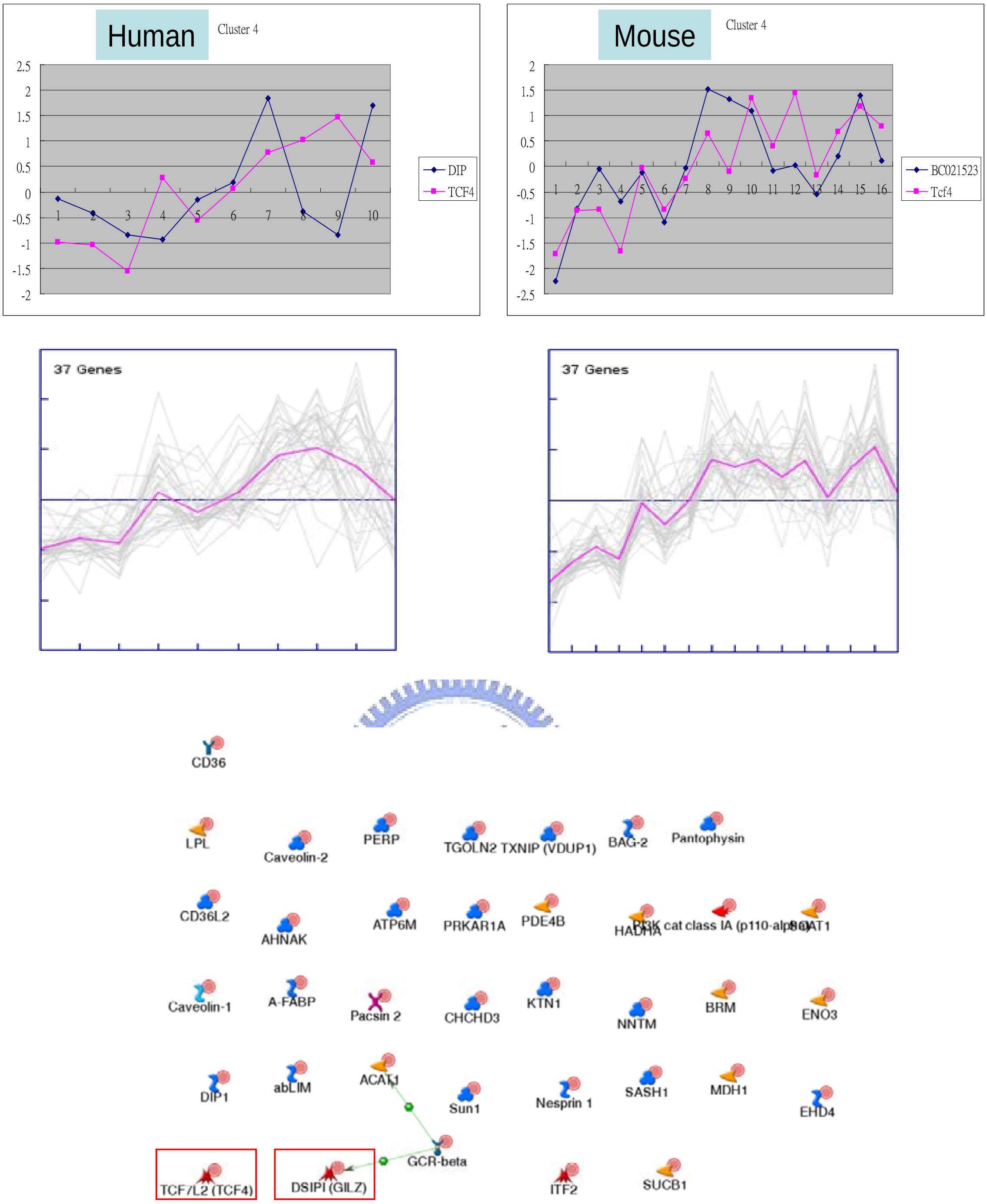


Figure 3.8 Transcription factors in cluster 4.

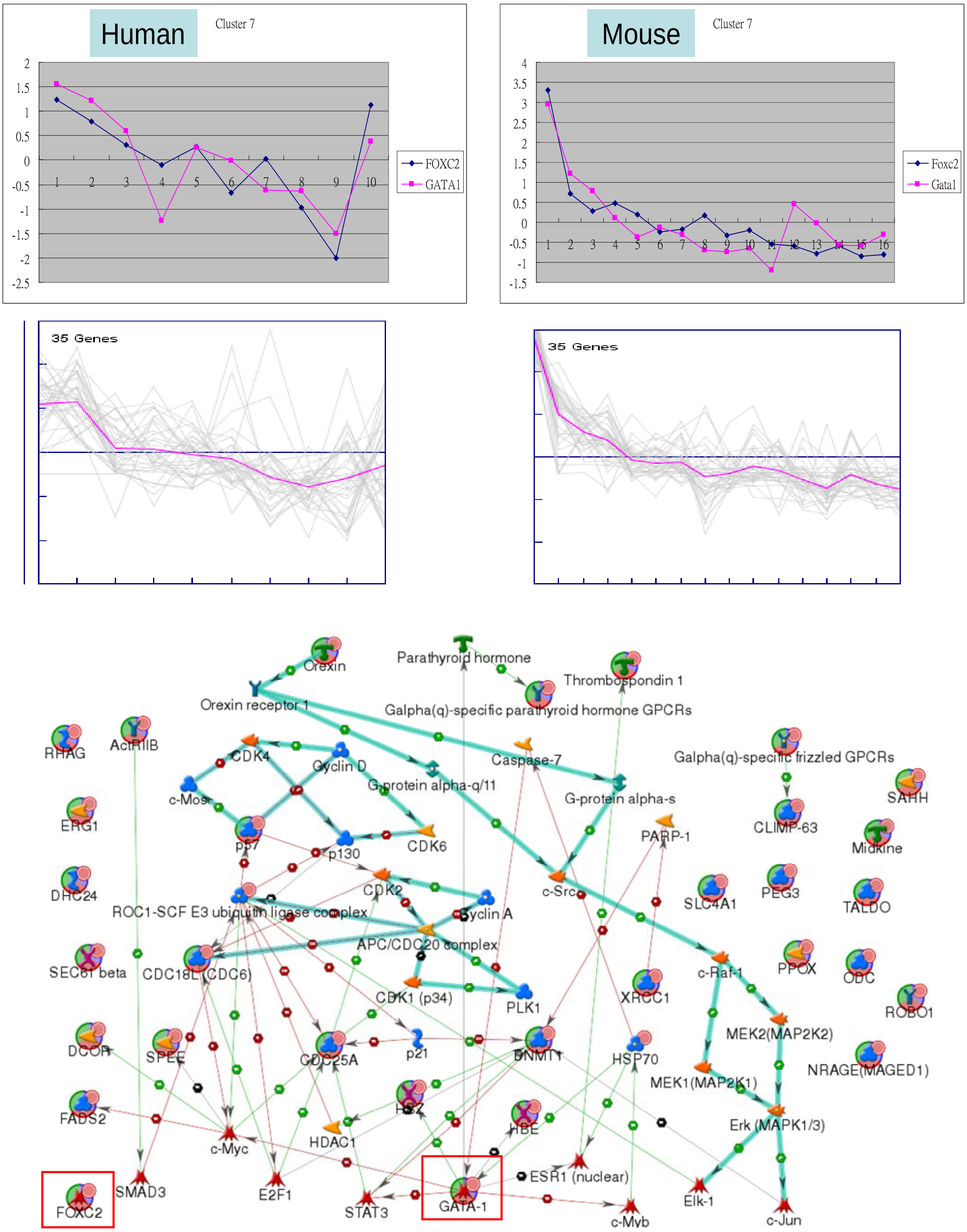


Figure 3.9 Transcription factors in cluster 7.

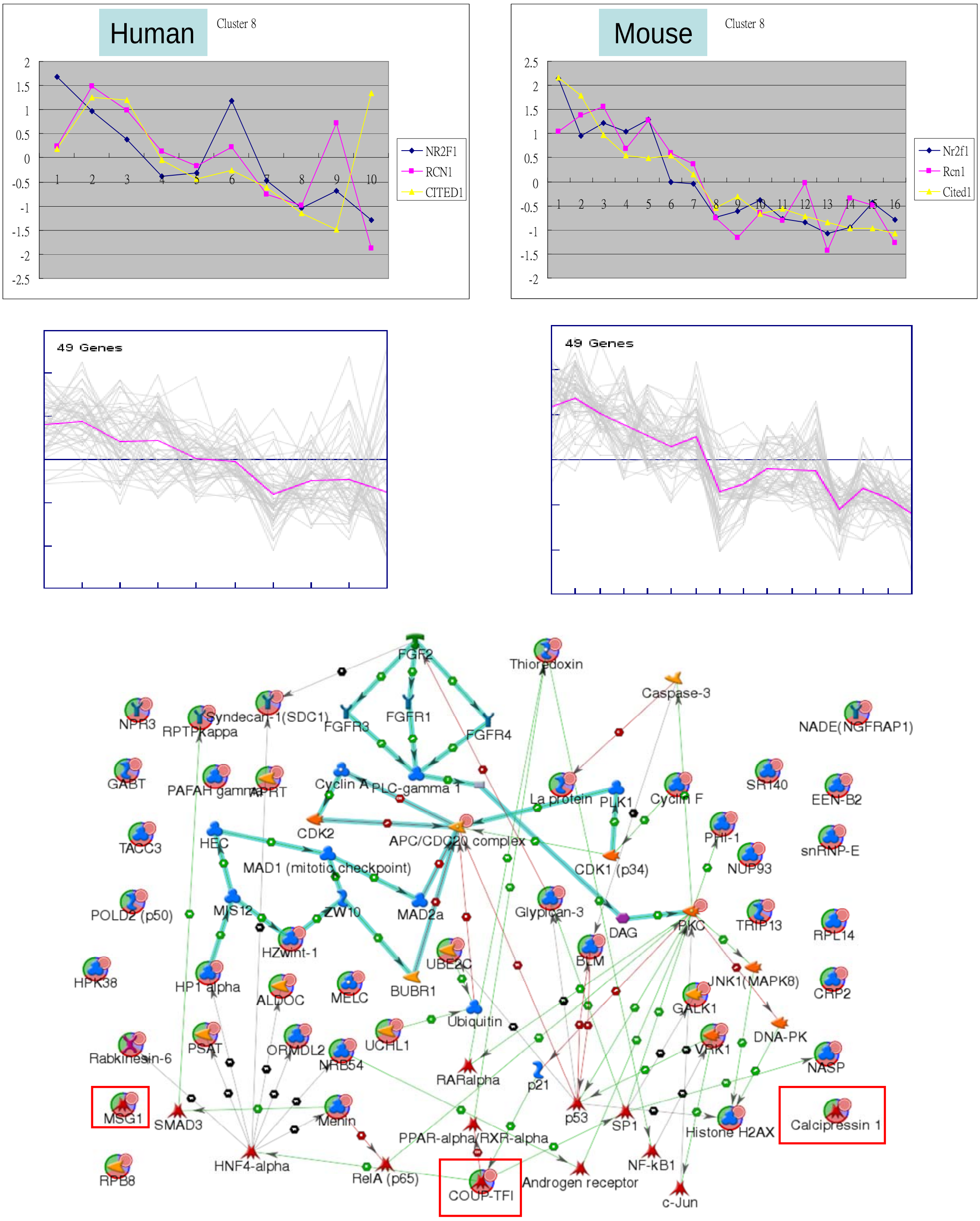


Figure 3.10 Transcription factors in cluster 8.

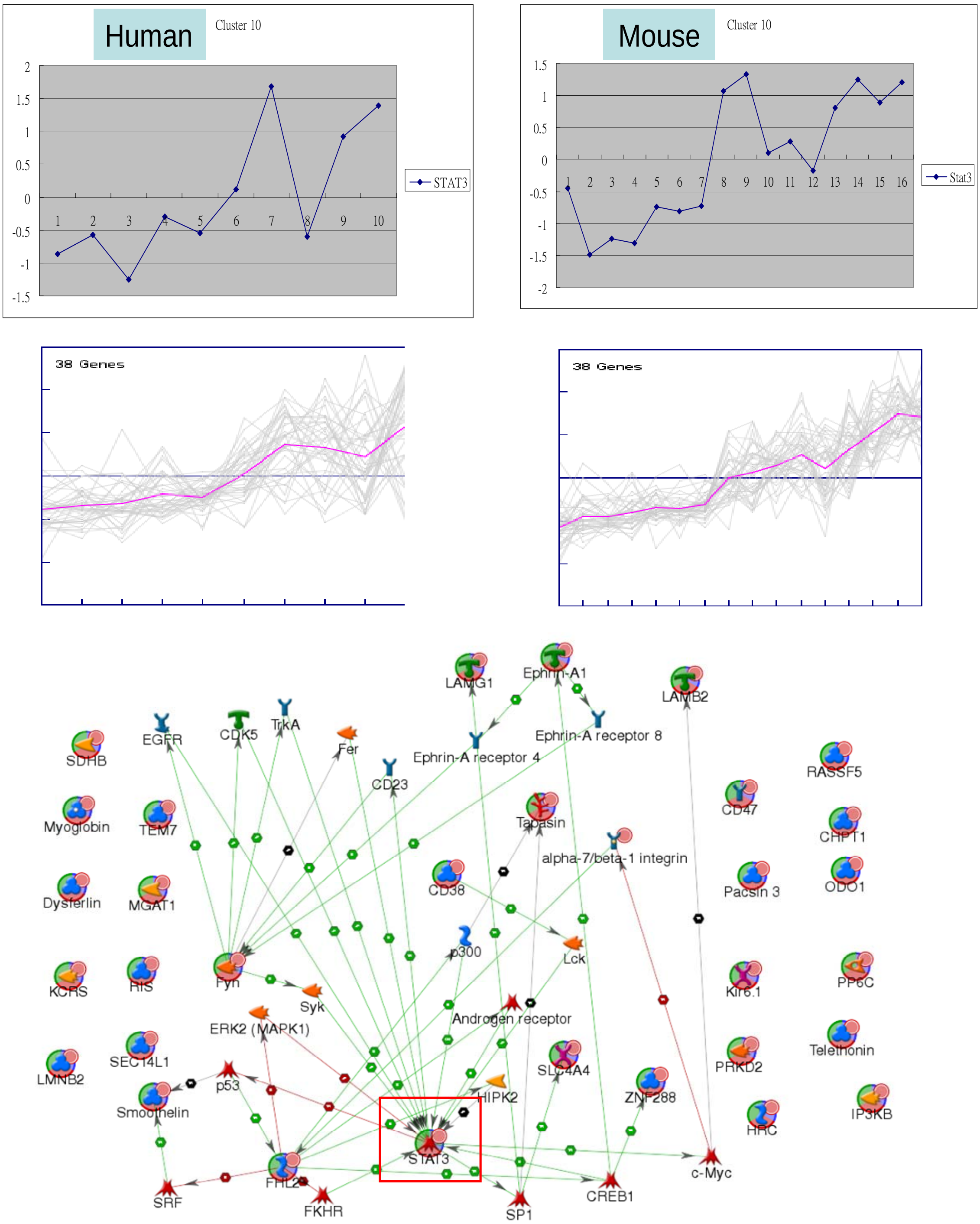


Figure 3.11 Transcription factors in cluster 10

3.7 Promoter Analysis of the Gene Groups

Based on the analysis of MetaCore, the regulatory network are built in each gene clusters. For example, two transcription factors, FOXC2 and GATA1, whose expression patterns are similar to other genes of cluster 7, regulate several target genes in cluster 7. However, there are several genes not regulated by FOXC2 and GATA1 based on the analysis of MetaCore. The genes which are not annotated that they are regulated by FOXC2 and GATA1 may be the targets of FOXC2 and GATA1. Therefore, the promoter sequences of genes which are not regulated by FOXC2 and GATA1 are used to scan whether the potential FOXC2 and GATA1 binding site on their promoter region or not. Four gene clusters which contain transcription factors are selected to analyze the transcription factor binding site by using the binding profile of TRANSFAC.

The analyzing flowchart of promoter analysis are illustrated in **Figure 3.12**, which containing gene clustering, promoter extraction, and TF binding site scanning. The genes which have similar expression patterns are clustering together by K-mean cluster method. The clustered genes are firstly analyzed by MetaCore for observing the transcription factor and regulatory network. On one hand, all genes other than transcription factor are selected to map the Ensembl gene ID and extract the promoter sequence which is defined as the region from upstream 2000 to downstream 200 of transcription start site (TSS). On the other hand, the transcription factor is mapped to TRANSFAC [21] factor ID and extracted the TF binding matrix. The TF binding matrix can be used by MATCH program to scan the TF binding sites on user input sequences with two important parameters, core similarity and matrix similarity. We set the core similarity to 100%, and the predicted binding sites on promoter sequences are graphically visualized, as shown in **Figure 3.13**.

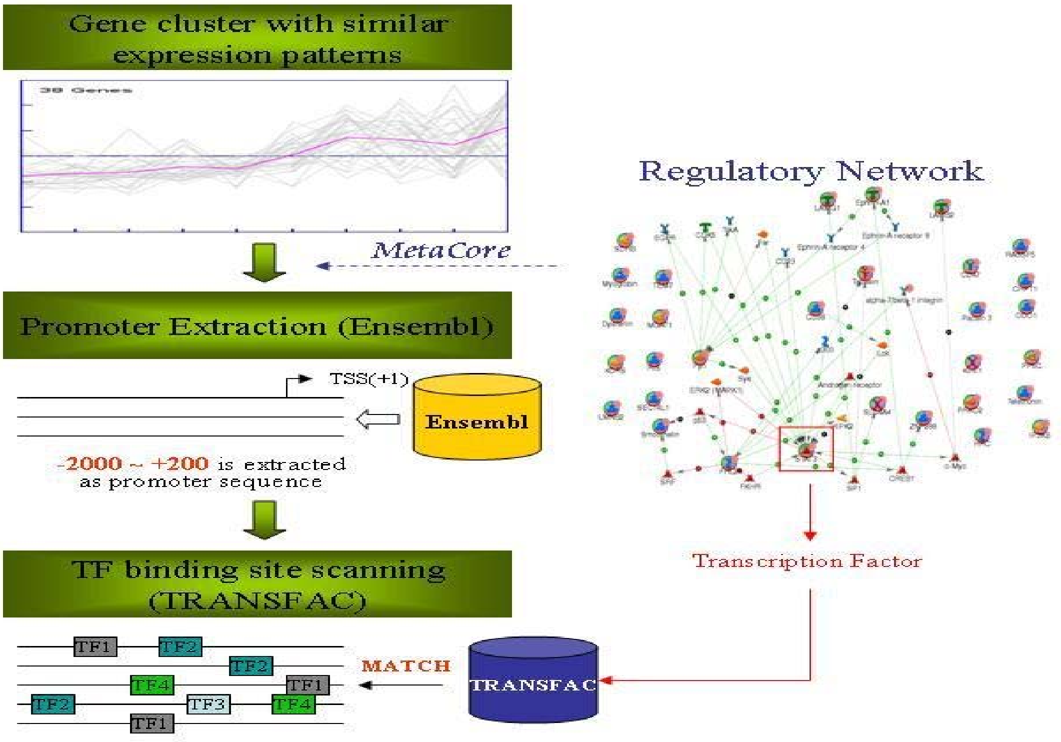

Gene cluster with similar expression patterns
39 Genes
MetaCore
Promoter Extraction (Ensembl)
TSS(+1)
Ensembl
-2000 ~ +200 is extracted as promoter sequence
TF binding site scanning (TRANSFAC)
TF1
TF2
TF2
TF4
TF1
TF2
TF3
TF4
TF1
MATCH
TRANSFAC
Regulatory Network
Transcription Factor

Figure 3.12 The analyzing flowchart of extracting promoter sequences and scanning TF binding site.

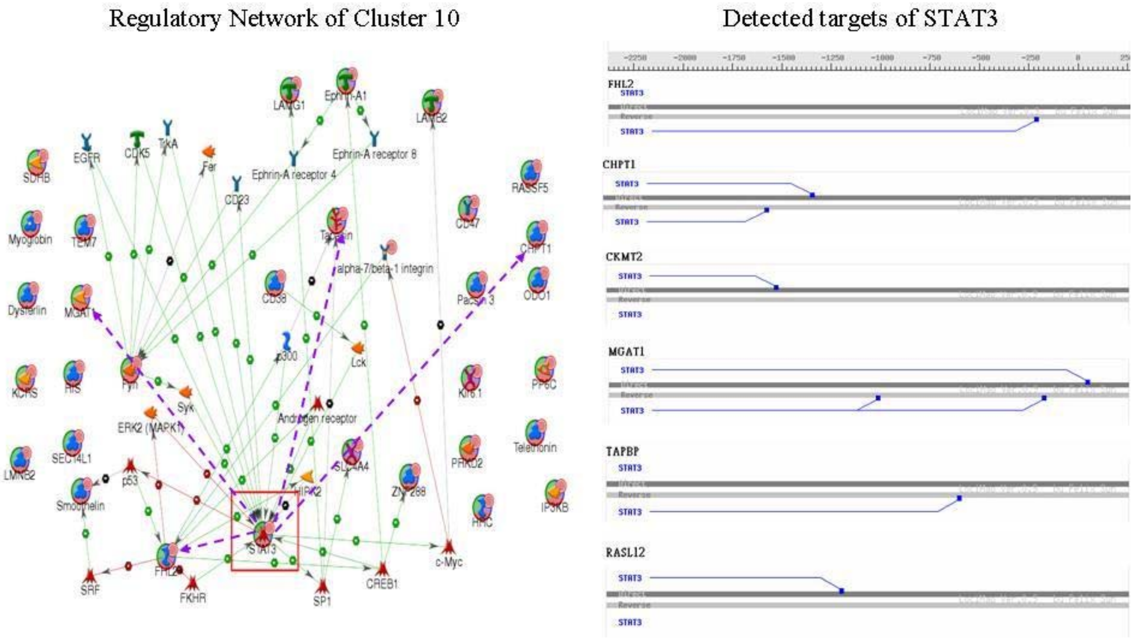


Figure 3.13 The detected targets of STAT3 transcription factor.

3.8 Validation of the discovery by referring to previous works

3.8.1 TGF and *Wnt* family

Activin/TGF-β and BMP-2/BMP-4 have distinct and reciprocal heart field mesoderm-inducing capacities that mimic the tissues in which they are expressed, the pregastrula hypoblast and anterior lateral endoderm, respectively[22]. Activin, TGF-β, and certain BMPs, which are members of the TGF-β superfamily, can mimic aspects of cardiogenesis, but none of these signaling peptides can induce the full range of activities elicited by the inducing tissues, nor do they show the capacity to convert noncardiogenic mesoderm toward a myocardial phenotype. The BMP type IA receptor called ALK3, along with TAK1 (mitogen-activated protein kinase kinase kinase) and Smad1, which are activated by BMP signaling, are coexpressed in the cardiogenic mesoderm [23, 24].

The biological pathway TGF, WNT and cytoskeletal remodeling and WNT signaling pathway are very significant in our 250 time-warped genes (see in **Table 3.5**). It is the validation of our results that these genes play key roles in heart development.

Table 3.5 Significant biological pathway of 250 Time-warped genes.

Map	Cell process	P-value	Genes

Propionate metabolism		3.93E-05	5	22
TGF, WNT and cytoskeletal remodeling	cell adhesion	9.33E-05	13	204
Chemokines and adhesion	cytokine and chemokine mediated signaling pathway, cell adhesion	3.45E-04	11	174
TCA		4.13E-04	4	20
Urea cycle		1.35E-03	4	27
Tryptophan metabolism		2.29E-03	4	31
Role of VDR in regulation of genes involved in osteoporosis	transcription	3.72E-03	5	57
WNT signaling pathway	response to extracellular stimulus	3.79E-03	6	82
Cytoskeleton remodeling	cell adhesion	4.96E-03	9	176
Prolactin receptor signaling	response to hormone stimulus, intracellular receptor-mediated signaling pathway	5.34E-03	5	62

3.8.2 GATA-4

The *GATA* gene family encodes transcription factors characterized by zinc-finger motifs required for DNA recognition, DNA binding, and transcription transcription activation [25]. Three members of the GATA family of transcription factors, *GATA4, 5,* and *6,* are expressed in the developing heart. *GATA5* is restricted to the endocardium while *GATA4* and *6* are expressed in the myocardium. The expression pattern of *GATA4* in the putative heart field encompasses that of *Nkx2.5,* but extends to a larger portion of the lateral plate mesoderm [26]. It has been proposed that combinatorial interaction among GATA factors or between GATA factors and other cofactors may differentially control various stages of cardiogenesis [27].

In cluster 7, there is a gene, GATA1, belong to the *GATA* gene family. GATA transcription factors play an important role in regulating the expression of many of the genes encoding myocardial contractile proteins, including cardiac troponin I, a gene that is expressed exclusively in cardiac myocytes [28, 29]; cardiac troponin C [30]; slow myosin heavy chain 3 [31]; and cardiac alpha actin[32, 33]. In addition, a number of other genes are responsive to GATA factors. These include early expression of *Nkx2.5* [34]; the atrial natriuretic factor [32, 35, 36]; a cardiac subtype of the muscarinic aceytcholine receptor [37]; and the sodium-calcium exchanger [38, 39]. In many cases, up-regulation of these genes requires the presence of other transcriptional partners such as serum response factor, MEF2C, or Nkx2.5.

3.9 Comparison to GEO Data

3.9.1 Human Data vs. GEO Mouse data

Our human data has 10 time points on heart embryo developmental stage; the GEO mouse data has 7 time points on the same condition. We applied the same time-warping method to these two datasets, and also selected 250 best time-warped genes. Finally, 62 genes were overlapped between the previous result and this result. In this analysis, the GEO data could be used to validate whether our human data is stable or not. For example, these 62 genes are very stable genes just because they are selected in two analyses. The cutoff value (250) can be adjust to bigger if we want to get more stable genes for further analysis. Using GEO data is a validation step and it makes the result more reliable. The expression profiles of these 63 genes are shown in **Figure 3.14**.

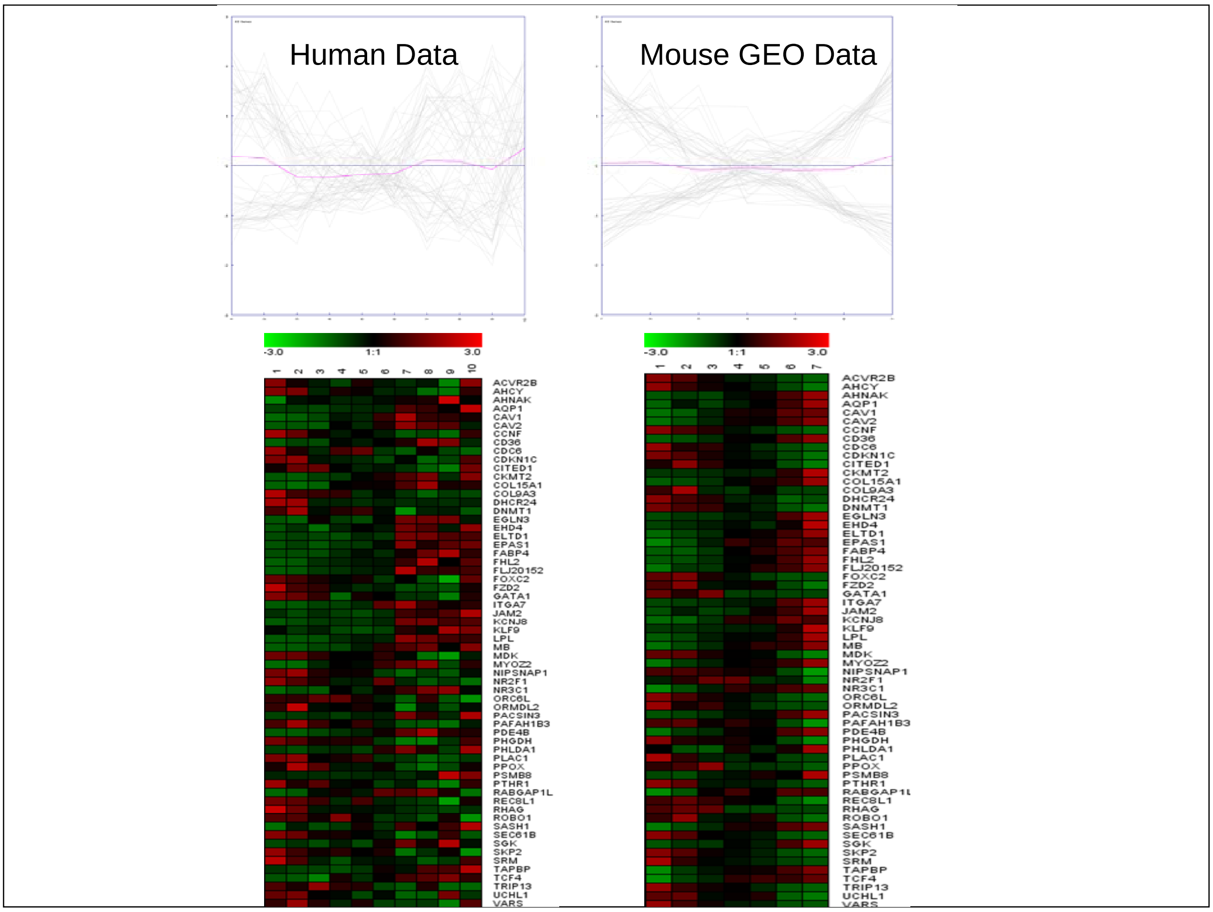


Figure 3.14 Expressions of 62 overlapped genes.

3.9.2 Mouse Data vs. GEO Mouse data

Our mouse data has 16 time points on heart developmental stage, among the 16 time points, 7 time points are on the embryonic development and 9 time points are on the fetal development. According to the GEO data is all from the embryonic development stage. We used our mouse data on the same condition, it means just seven time points was used in our mouse data. We applied the same time-warping method to these two datasets, and also selected 250 best time-warped genes. Finally, 37 genes were overlapped between our original result and this result. In this analysis, the GEO data could be used to validate whether our mouse data is stable or not. For example, these 65 genes are very stable genes just because they are selected in two analyses. The cutoff value (250) can be adjust to bigger if we want to get more stable genes for further analysis. Using GEO data is a validation step and it makes the result more reliable. The expression profiles of these 37 genes are shown in **Figure 3.15**.

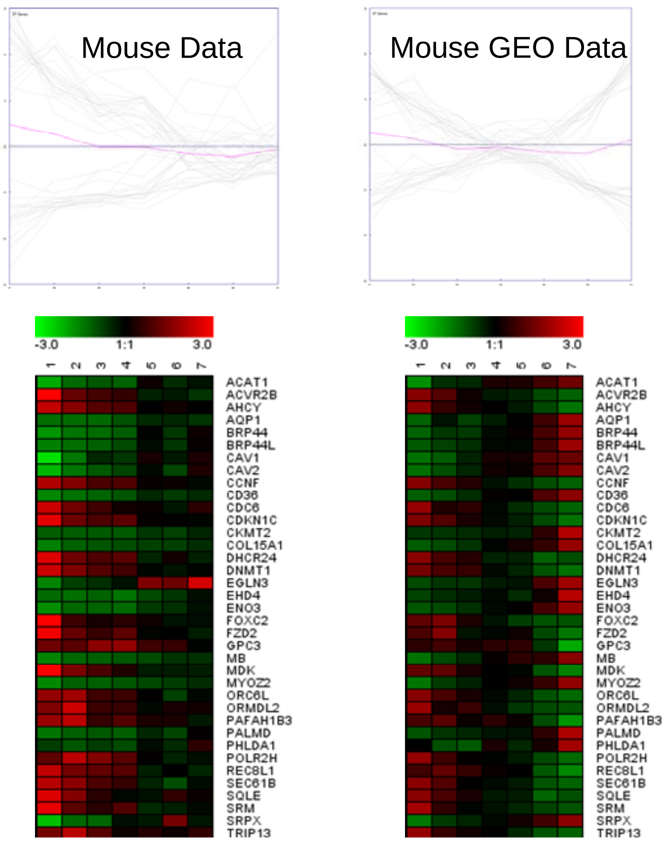


Figure 3.15 Expressions of 37 overlapped genes.

Chpater 4 Discussions

4.1 Study limitations

In most of our experimental procedures, however, we have to grind the tissue, extract RNA, and analyze the changes of each gene along with development age. As a consequence, the acute limitation of the results derived from this study is the lack of spatial patterning of each gene, for instance, in two dimensions or three dimensions. Nevertheless, results of this study will provide an ontogeny map of gene expression profiles, from which we can identify groups of temporal and spatial information to facilitate our understanding of the human developmental biology.

Furthermore, since the gene expression profiles in heart of the fetus have been identified to be similar to those in corresponding types of cancer[40, 41] and those of failing heart [42, 43] or dysfunctional heart, knowledge advances in the human early development, at the transcriptional level, will cast insights not only into the molecular mechanisms of human chromosomal anomalies but also into that of dysfunction and regenerative diseases.

4.2 Prospective works

4.2.1 Analyzing Gene Expression Profiles of Human and Mouse among Different Tissues during Embryonic Development

In this study, we only focus on the fetal age-specific gene expression profiles in one tissue (heart). In order to get more understanding of gene expressions of other tissues, we have to produce more microarray data in other tissues such as brain, lung, liver, kidney, and muscle-----etc. We expect that results from the study we propose here will provide the complete data, at the transcriptional level in different tissues, about the fetal developmental equivalence between the human and the mouse.

As soon as we verify the temporal changes using multiple mice specimens at each time point through the aforementioned comparative genomic study, we can depict the development-specific gene expression profiles of each tissue among human and mouse fetuses. This information will provide an invaluable developmental biology database of these tissues.

The database will serve as an indispensable reference for analyzing the changes of gene expression in the age-matched abnormal fetuses, such as in various types of trisomy and contiguous chromosomal syndromes.

4.2.2 Determination of Abnormal Genes in Development

Upon the confirmation of the human age-specific development gene expression profiles, we can perform DNA chips to analyze the gene expression profiles in different tissues to detect possible disease-related changes in gene expression profiles. Specifically, we will focus on those genes that have been mapped to corresponding abnormal chromosomes. In order to gain a better understanding on the abnormal fetuses, we may produce the following steps：(1) identification of dysfunctional expression profiles in target tissues of aneuploid fetuses, (2) determination of tissue-specificity of gene expression in contiguous chromosomal deletion/amplification syndrome, (3) validation of the role of candidate genes during development using gene knock-out mouse models, and (4) cellular and molecular functional analysis of the genes, which exhibit tissue-specific importance during fetal development, in the corresponding cancer cell lines.

4.2.3 Validation of the role of candidate genes during development using conditional gene knock-out mouse model

The ultimate confirmation of the role for a gene in fetal development is to create a mouse model with knocking-out (KO) of the orthologous gene, and follow the embryogenesis of fetal mice. If KO the gene of interest gene causes fetal lethality, we should pay more attention to the earlier embryonic age when the fetal demise occurs and to the detection of any associated developmental disorder. If the gene KO does not cause fetal lethality, we will carefully follow the change in litter size, the ratio between both sexes in the littermates, the growth pattern in terms of body weight gain in every week, sexual maturity, fertility, and whether the KO mice develop any natural diseases that may be common in C57Bl/6J mice earlier than normal controls, etc. In this scenario, it is still worthwhile to perform the systematic analyses of gene expression profiles in developing KO fetuses, and to compare those profiles with the temporal change of gene expression profiles in normal controls.

4.2.4 Multiple Alignment and Local Alignment

Because the limitation of the program, genewarp, only two datasets could be used to do time-warping. For this reason, only two species or two groups of genes can map together with this algorithm. We want to develop a tool which can provide user to do more than two datasets dynamic time-warping. Then, we can apply this method to do more comprehensive analyses between more species and more tissues….etc. For example, we can implement our method in human, mouse and rat.

In our research, global alignment is used with all of the data. As we know, time series data has a problem. How to sample the time points in the development stage? Is it enough? or too much? Global alignment utilizes all the time points given in the dataset. But some important genes just expressed in some period of time in embryogenesis. At this time, local alignment becomes more suitable for the analysis. In the further, we hope to find some important developmental genes between human and mouse, and see how they map in the period among the time points be given by using local alignment.

4.3 Conclusion

In conclusion, after working on the high-throughput functional genomics using DNA microarray technology, the most important thing is：Whatever gene that is discovered by the high-throughput screening or profiling methods should be carefully followed up with solid and thorough verification using conventional cell and molecular biological techniques.

Bibliography

1. Ton, C., et al., *Construction of a zebrafish cDNA microarray: gene expression profiling of the zebrafish during development.* Biochem Biophys Res Commun, 2002. **296**(5): p. 1134-42.
2. Lo, J., et al., *15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis.* Genome Res, 2003. **13**(3): p. 455-66.
3. Furlong, E.E., et al., *Patterns of gene expression during Drosophila mesoderm development.* Science, 2001. **293**(5535): p. 1629-33.
4. Small, C.L., et al., *Profiling gene expression during the differentiation and development of the murine embryonic gonad.* Biol Reprod, 2005. **72**(2): p. 492-501.
5. Yu, Y., et al., *Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs.* Genome Res, 2001. **11**(8): p. 1392-403.
6. Mody, M., et al., *Genome-wide gene expression profiles of the developing mouse hippocampus.* Proc Natl Acad Sci U S A, 2001. **98**(15): p. 8862-7.
7. Hamatani, T., et al., *Dynamics of global gene expression changes during mouse preimplantation development.* Dev Cell, 2004. **6**(1): p. 117-31.
8. Zaffran, S. and M. Frasch, *Early signals in cardiac development.* Circ Res, 2002. **91**(6): p. 457-69.

9. Kirby, M.L., *Molecular embryogenesis of the heart.* Pediatr Dev Pathol, 2002. **5**(6): p. 516-43.
10. Wang, T.H., et al., *Paclitaxel (Taxol) upregulates expression of functional interleukin-6 in human ovarian cancer cells through multiple signaling pathways.* Oncogene, 2006. **25**(35): p. 4857-66.
11. Chao, A., et al., *Molecular characterization of adenocarcinoma and squamous carcinoma of the uterine cervix using microarray analysis of gene expression.* Int J Cancer, 2006. **119**(1): p. 91-8.
12. Lee, Y.S., et al., *Molecular signature of clinical severity in recovering patients with severe acute respiratory syndrome coronavirus (SARS-CoV).* BMC Genomics, 2005. **6**: p. 132.
13. Wang, T.H., et al., *Establishment of cDNA microarray analysis at the Genomic Medicine Research Core Laboratory (GMRCL) of Chang Gung Memorial Hospital.* Chang Gung Med J, 2004. **27**: p. 243-60.
14. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
15. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.
16. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2005. **33**(Database issue): p. D39-45.
17. 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.
18. Aach, J. and G.M. Church, *Aligning gene expression time series with time warping algorithms.* Bioinformatics, 2001. **17**(6): p. 495-508.
19. Kruskal, J.B. and M. Liberman, *The symmetric time-warping problem: from continous to discrete. In Sankoff,D. and Kruskal,J. (eds),.* Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison 1999: CSLI Publications, Stanford,. pp. 125-161.
20. Al-Shahrour, F., et al., *BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W460-4.
21. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic Acids Res, 2003. **31**(1): p. 374-8.
22. Ladd, A.N., T.A. Yatskievych, and P.B. Antin, *Regulation of avian cardiac myogenesis by activin/TGFbeta and bone morphogenetic proteins.* Dev Biol, 1998. **204**(2): p. 407-19.
23. Delot, E.C., et al., *BMP signaling is required for septation of the outflow tract of the mammalian heart.* Development, 2003. **130**(1): p. 209-20.
24. Monzen, K., et al., *Bone morphogenetic proteins induce cardiomyocyte differentiation through the mitogen-activated protein kinase kinase kinase TAK1 and cardiac transcription factors Csx/Nkx-2.5 and GATA-4.* Mol Cell Biol, 1999. **19**(10): p. 7096-105.
25. Weiss, M.J. and S.H. Orkin, *GATA transcription factors: key regulators of hematopoiesis.* Exp Hematol, 1995. **23**(2): p. 99-107.
26. Serbedzija, G.N., J.N. Chen, and M.C. Fishman, *Regulation in the heart field of zebrafish.* Development, 1998. **125**(6): p. 1095-101.
27. Charron, F. and M. Nemer, *GATA transcription factors and cardiac development.* Semin Cell Dev Biol, 1999. **10**(1): p. 85-91.
28. Di Lisi, R., et al., *Combinatorial cis-acting elements control tissue-specific activation of the cardiac troponin I gene in vitro and in vivo.* J Biol Chem, 1998. **273**(39): p. 25371-80.
29. Murphy, A.M., et al., *Regulation of the rat cardiac troponin I gene by the transcription factor GATA-4.* Biochem J, 1997. **322 (Pt 2)**: p. 393-401.
30. Ip, H.S., et al., *The GATA-4 transcription factor transactivates the cardiac muscle-specific troponin C promoter-enhancer in nonmuscle cells.* Mol Cell Biol, 1994. **14**(11): p. 7517-26.
31. Wang, G.F., et al., *A positive GATA element and a negative vitamin D receptor-like element control atrial chamber-specific expression of a slow myosin heavy-chain gene during cardiac morphogenesis.* Mol Cell Biol, 1998. **18**(10): p. 6023-34.
32. Sepulveda, J.L., et al., *GATA-4 and Nkx-2.5 coactivate Nkx-2 DNA binding targets: role for regulating early cardiac gene expression.* Mol Cell Biol, 1998. **18**(6): p. 3405-15.
33. Durocher, D., et al., *The cardiac transcription factors Nkx2-5 and GATA-4 are mutual cofactors.* Embo J, 1997. **16**(18): p. 5687-96.
34. Searcy, R.D. and K.E. Yutzey, *Analysis of Hox gene expression during early avian heart development.* Dev Dyn, 1998. **213**(1): p. 82-91.
35. Shiojima, I., et al., *Context-dependent transcriptional cooperation mediated by cardiac transcription factors Csx/Nkx-2.5 and GATA-4.* J Biol Chem, 1999. **274**(12): p. 8231-9.

36. Lee, Y., et al., *The cardiac tissue-restricted homeobox protein Csx/Nkx2.5 physically associates with the zinc finger protein GATA4 and cooperatively activates atrial natriuretic factor gene expression.* Mol Cell Biol, 1998. **18**(6): p. 3120-9.

37. Rosoff, M.L. and N.M. Nathanson, *GATA factor-dependent regulation of cardiac m2 muscarinic acetylcholine gene transcription.* J Biol Chem, 1998. **273**(15): p. 9124-9.

38. Nicholas, S.B. and K.D. Philipson, *Cardiac expression of the Na(+)/Ca(2+) exchanger NCX1 is GATA factor dependent.* Am J Physiol, 1999. **277**(1 Pt 2): p. H324-30.

39. Cheng, G., et al., *The role of GATA, CArG, E-box, and a novel element in the regulation of cardiac expression of the Na+-Ca2+ exchanger gene.* J Biol Chem, 1999. **274**(18): p. 12819-26.

40. Schaaf, G.J., et al., *Full transcriptome analysis of rhabdomyosarcoma, normal, and fetal skeletal muscle: statistical comparison of multiple SAGE libraries.* Faseb J, 2005. **19**(3): p. 404-6.

41. Lechner, J.F., et al., *Human lung cancer cells and tissues partially recapitulate the homeobox gene expression profile of embryonic lung.* Lung Cancer, 2002. **37**(1): p. 41-7.

42. Razeghi, P., et al., *Metabolic gene expression in fetal and failing human heart.* Circulation, 2001. **104**(24): p. 2923-31.

43. Depre, C., et al., *Unloaded heart in vivo replicates fetal gene expression of cardiac hypertrophy.* Nat Med, 1998. **4**(11): p. 1269-75.

Appendix A

Cluster	Human gene	Mouse gene	Score	Chromosome	Ensembl Gene ID	Description
1 (11 genes)	FZD2	Fzd2	2.35004	chr17q21.1	ENSG00000180340	frizzled homolog 2 (Drosophila)
	MB	Mb	2.37105	chr22q13.1	ENSG00000198125	myoglobin
	FHL2	Fhl2	2.49196	chr2q12-q14	ENSG00000115641	four and a half LIM domains 2
	CBX5	Cbx5	2.51184	chr12q13.13	ENSG00000094916	chromobox homolog 5 (HP1 alpha homolog, Drosophila)
	CKMT2	Ckmt2	2.62403	chr5q13.3	ENSG00000131730	creatine kinase, mitochondrial 2 (sarcomeric)
	LPL	Lpl	2.64993	chr8p22	---	lipoprotein lipase
	TGOLN2	Tgoln1	2.65776	chr2p11.2	ENSG00000152291	trans-golgi network protein 2
	PAFAH1B3	Pafah1b3	2.73228	chr19q13.1	ENSG00000079462	platelet-activating factor acetylhydrolase, isoform Ib, gamma subunit 29kDa
	B2M	B2m	2.75425	chr15q21-q22.2	ENSG00000166710	beta-2-microglobulin
	COL15A1	Col15a1	2.76479	chr9q21-q22	ENSG00000204291	collagen, type XV, alpha 1
	RHAG	Rhag	2.77558	chr6p21.1-p11	ENSG00000112077	Rh-associated glycoprotein
2 (14 genes)	MAGED1	Maged1	2.78243	chrXp11.23	ENSG00000179222	melanoma antigen family D, 1
	NIPSNAP1	Nipsnap1	2.89711	chr22q12.2	ENSG00000184117	nipsnap homolog 1 (C. elegans)
	JAM2	Jam2	2.92219	chr21q21.2	ENSG00000154721	junctional adhesion molecule 2
	SMTN	Smtn	2.95343	chr22q12.2	ENSG00000183963	smoothelin
	HLA-DRA	H2-Ea	3.03231	chr6p21.3	ENSG00000204287 /// ENSG00000206243 /// ENSG00000206308	major histocompatibility complex, class II, DR alpha
	SRM	Srm	3.03747	chr1p36-p22	ENSG00000116649	spermidine synthase
	CSRP2	Csrp2	3.11019	chr12q21.1	ENSG00000175183	cysteine and glycine-rich protein 2
	ABAT	Abat	3.12856	chr16p13.2	ENSG00000183044	4-aminobutyrate aminotransferase

	APRT	Aprt	3.14577	chr16q24	ENSG00000198931	adenine phosphoribosyltransferase
	AHCY	Ahcy	3.17586	chr20cen-q13.1	ENSG00000101444	S-adenosylhomocysteine hydrolase
	CAV2	Cav2	3.17918	chr7q31.1	---	Caveolin 2
	HBE1	Hbb-y	3.18354	chr11p15.5	ENSG00000196565	hemoglobin, epsilon 1 /// hemoglobin, epsilon 1
	C6orf108	BC048355	3.20422	chr6p21.1	ENSG00000112667	chromosome 6 open reading frame 108
	SYNE1	Syne1	3.21439	chr6q25	---	spectrin repeat containing, nuclear envelope 1
3 (16 genes)	NR2F1	Nr2f1	3.23716	chr5q14	ENSG00000175745	nuclear receptor subfamily 2, group F, member 1
	VSNL1	Vsnl1	3.23832	chr2p24.3	ENSG00000163032	visinin-like 1
	KIF20A	Kif20a	3.24549	chr5q31	ENSG00000112984	kinesin family member 20A
	DUSP1	Dusp1	3.25684	chr5q34	ENSG00000120129	dual specificity phosphatase 1
	ELTD1	Eltd1	3.26212	chr1p33-p32	ENSG00000162618	EGF, latrophilin and seven transmembrane domain containing 1
	ITPKB	Itpkb	3.2697	chr1q42.13	ENSG00000143772	inositol 1,4,5-trisphosphate 3-kinase B
	RNF8	Rnf8	3.27455	chr6p21.3	ENSG00000112130	ring finger protein 8
	AKAP1	Akap1	3.28084	chr17q21-q23	ENSG00000121057	A kinase (PRKA) anchor protein 1
	CHCHD3	Chchd3	3.2875	chr7q32.3-q33	ENSG00000106554	coiled-coil-helix-coiled-coil-helix domain containing 3
	CDKN1C	Cdkn1c	3.31253	chr11p15.5	ENSG00000129757	cyclin-dependent kinase inhibitor 1C (p57, Kip2)
	TCF4	Tcf4	3.32639	chr18q21.1	ENSG00000196628	transcription factor 4
	PEG3	Peg3	3.32742	chr19q13.4	ENSG00000198300	paternally expressed 3
	EFNA1	Efna1	3.33508	chr1q21-q22	ENSG00000169242	ephrin-A1
	IL7R	Il7r	3.34235	chr5p13	ENSG00000168685	interleukin 7 receptor /// interleukin 7 receptor
	LMOD1	Lmod1	3.3425	chr1q32	ENSG00000163431	leiomodin 1 (smooth muscle)
	NNT	Nnt	3.34296	chr5p13.1-5cen	---	Nicotinamide nucleotide transhydrogenase

4 (37 genes)	PIP5K1B	Pip5k1a	3.34681	chr9q13	---	phosphatidylinositol-4-phosphate 5-kinase, type I, beta
	CA5B	Car5b	3.34887	chrXp21.1	---	carbonic anhydrase VB, mitochondrial
	BAG2	Bag2	3.37544	chr6p12.3-p11.2	ENSG00000112208	BCL2-associated athanogene 2
	EBI2	Ebi2	3.37733	chr13q32.3	ENSG00000169508	Epstein-Barr virus induced gene 2 (lymphocyte-specific G protein-coupled receptor)
	NUP93	Nup93	3.38446	chr16q13	ENSG00000102900	nucleoporin 93kDa
	CCNF	Ccnf	3.39695	chr16p13.3	ENSG00000162063	cyclin F
	TMEM59	Tmem59	3.40408	chr1p36-p31	ENSG00000116209	transmembrane protein 59
	PTHR1	Pthr1	3.40624	chr3p22-p21.1	ENSG00000160801	parathyroid hormone receptor 1
	TACC3	Tacc3	3.40738	chr4p16.3	ENSG00000013810	transforming, acidic coiled-coil containing protein 3
	SH3GL3	Sh3gl3	3.40802	chr15q24	---	SH3-domain GRB2-like 3
	EPAS1	Epas1	3.41156	chr2p21-p16	ENSG00000116016	endothelial PAS domain protein 1
	REC8L1	Rec8L1	3.41256	chr14q11.2-q12	ENSG00000100918	REC8-like 1 (yeast)
	KCNJ8	Kcnj8	3.41825	chr12p11.23	ENSG00000121361	potassium inwardly-rectifying channel, subfamily J, member 8
	FABP4	Fabp4	3.42139	chr8q21	ENSG00000170323	fatty acid binding protein 4, adipocyte
	GPM6B	Gpm6b	3.42442	chrXp22.2	ENSG00000046653	glycoprotein M6B
	HBZ	Hba-x	3.45424	chr16p13.3	ENSG00000101442	Hemoglobin, zeta
	CD36	Cd36	3.45645	chr7q11.2	---	CD36 molecule (thrombospondin receptor)
	NR3C1	Nr3c1	3.46739	chr5q31.3	ENSG00000113580	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
	PPP1R14B	Ppp1r14b	3.48588	chr11q13	ENSG00000173457	protein phosphatase 1, regulatory (inhibitor) subunit 14B
	FLJ22662	1100001H23Rik	3.48925	chr12p13.1	ENSG00000121316	hypothetical protein FLJ22662

	PDE4B	Pde4b	3.50097	chr1p31	---	Phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila)
	KIAA0141	0610009O20Rik	3.51167	chr5q31.3	ENSG00000081791	KIAA0141
	PACSIN3	Pacsin3	3.52113	chr11p12-p11.12	ENSG00000165912	protein kinase C and casein kinase substrate in neurons 3
	PPARGC1A	Ppargc1a	3.52844	chr4p15.1	ENSG00000109819	peroxisome proliferator-activated receptor gamma, coactivator 1 alpha
	DHCR24	Dhcr24	3.53595	chr1p33-p31.1	ENSG00000116133	24-dehydrocholesterol reductase
	CYSLTR2	Cysltr2	3.54782	chr13q14.12-q21.1	ENSG00000152207	cysteinyl leukotriene receptor 2
	BZRPL1	Bzrpl1	3.56339	chr6p21.1	ENSG00000112212	benzodiazapine receptor (peripheral)-like 1
	ZWINT	Zwint	3.5696	chr10q21-q22	ENSG00000122952	ZW10 interactor
	ZBTB20	Zbtb20	3.57055	chr3q13.2	---	zinc finger and BTB domain containing 20
	S100B	S100b	3.57258	chr21q22.3	ENSG00000160307	S100 calcium binding protein B
	DYSF	Dysf	3.57496	chr2p13.3-p13.1	ENSG00000135636	dysferlin, limb girdle muscular dystrophy 2B (autosomal recessive)
	SDC1	Sdc1	3.5805	chr2p24.1	ENSG00000115884	syndecan 1
	HIST1H2BD	Hist1h2bp	3.58233	chr6p21.3	ENSG00000158373	histone cluster 1, H2bd
	CYP1B1	Cyp1b1	3.58453	chr2p21	ENSG00000138061	cytochrome P450, family 1, subfamily B, polypeptide 1
	TAPBP	Tapbp	3.5885	chr6p21.3	ENSG00000112493	TAP binding protein (tapasin)
	MYL1	Myl1	3.59016	chr2q33-q34	ENSG00000168530	myosin, light chain 1, alkali; skeletal, fast
	MELK	Melk	3.59676	chr9p13.2	ENSG00000165304	maternal embryonic leucine zipper kinase
5 (3 genes)	ORMDL2	Ormdl2	3.60337	chr12q13.2	ENSG00000123353	ORM1-like 2 (S. cerevisiae)
	ITGB2	Itgb2	3.60614	chr21q22.3	ENSG00000160255	integrin, beta 2 (complement component 3 receptor 3 and 4 subunit)
	BRP44	Brp44	3.60802	chr1q24	ENSG00000143158	brain protein 44

6 (22 genes)	RASL11B	Rasl11b	3.61041	chr4q12	ENSG00000128045	RAS-like, family 11, member B
	ROBO1	Robo1	3.61483	chr3p12	ENSG00000169855	roundabout, axon guidance receptor, homolog 1 (Drosophila)
	FYN	Fyn	3.61523	chr6q21	---	FYN oncogene related to SRC, FGR, YES
	MDK	Mdk	3.61841	chr11p11.2	ENSG00000110492	midkine (neurite growth-promoting factor 2)
	FZD7	Fzd7	3.62471	chr2q33	ENSG00000155760	frizzled homolog 7 (Drosophila)
	DARS	Dars	3.62581	chr2q21.3	ENSG00000115866	aspartyl-tRNA synthetase
	ATP6V1D	Atp6v1d	3.63456	chr14q23-q24.2	---	ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D
	H2AFX	H2afx	3.6369	chr11q23.2-q23.3	---	H2A histone family, member X
	SASH1	Sash1	3.63903	chr6q24.3	ENSG00000111961	SAM and SH3 domain containing 1
	ACVR2B	Acvr2b	3.65439	chr3p22	ENSG00000114739	activin A receptor, type IIB
	HSPB8	Hspb8	3.65606	chr12q24.23	ENSG00000152137	heat shock 22kDa protein 8
	FXYD6	Fxyd6	3.6581	chr11q23.3	ENSG00000137726	FXYD domain containing ion transport regulator 6
	FADS2	Fads2	3.66594	chr11q12-q13.1	ENSG00000134824	fatty acid desaturase 2
	CD47	Cd47	3.66923	chr3q13.1-q13.2	ENSG00000196776	CD47 molecule
	HIGD1A	Higd1a	3.66961	chr3p22.1	ENSG00000181061	HIG1 domain family, member 1A
	DIP	BC021523	3.68572	chr22q13.31	ENSG00000075240	death-inducing-protein
	SNRPE	Snrpe	3.68966	chr1q32	ENSG00000182004	small nuclear ribonucleoprotein polypeptide E
	AHNAK	Ahnak	3.70083	chr11q12.2	---	AHNAK nucleoprotein (desmoyokin)
	CLEC3B	Clec3b	3.70897	chr3p22-p21.3	ENSG00000163815	C-type lectin domain family 3, member B
	UNC84A	Unc84a	3.71529	chr7p22.3	ENSG00000164828	unc-84 homolog A (C. elegans)
	TNIP2	Tnip2	3.71928	chr4p16.3	ENSG00000168884	TNFAIP3 interacting protein 2
	FOXC2	Foxc2	3.72093	chr16q22-16q24	ENSG00000176692	forkhead box C2 (MFH-1, mesenchyme forkhead 1)
7 (35 genes)	PLAC1	Plac1	3.72157	chrXq26	ENSG00000170965	placenta-specific 1

BRP44L	Brp44l	3.72977	chr6q27	ENSG00000060762	brain protein 44-like
COL9A3	Col9a3	3.73265	chr20q13.3	ENSG00000092758	collagen, type IX, alpha 3
POLR2H	Polr2h	3.7444	chr3q28	ENSG00000163882	polymerase (RNA) II (DNA directed) polypeptide H
PIK3CA	Pik3ca	3.74511	chr3q26.3	---	Phosphoinositide-3-kinase, catalytic, alpha polypeptide
AP2S1	Ap2s1	3.74648	chr19q13.2-q13.3	ENSG00000042753	adaptor-related protein complex 2, sigma 1 subunit /// adaptor-related protein complex 2, sigma 1 subunit
UBE2C	Ube2c	3.74671			
TRIP13	Trip13	3.74956	chr5p15.33	ENSG00000071539	thyroid hormone receptor interactor 13
MRLC2	Mylc2b	3.75306	chr18p11.31	---	myosin regulatory light chain MRLC2
SKP2	Skp2	3.75813	chr5p13	ENSG00000145604	S-phase kinase-associated protein 2 (p45)
CDC20	Cdc20	3.75973	chr1p34.1	ENSG00000117399	cell division cycle 20 homolog (S. cerevisiae)
FLJ20152	1810015C04Rik	3.76006	chr5p15.1	ENSG00000154153	hypothetical protein FLJ20152
PERP	Perp	3.76096	chr6q24	ENSG00000112378	PERP, TP53 apoptosis effector
PSMB8	Psmb8	3.76747	chr6p21.3	ENSG00000204264 /// ENSG00000206234 /// ENSG00000206298	proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional peptidase 7)
RASL12	Rasl12	3.77026	chr15q11.2-q22.33	ENSG00000103710	RAS-like, family 12
SSB	Ssb	3.77204	chr2q31.1	ENSG00000138385	Sjogren syndrome antigen B (autoantigen La)
ALDOC	Aldoc	3.7721	chr17cen-q12	ENSG00000109107	aldolase C, fructose-bisphosphate
PALMD	Palmd	3.77349	chr1p22-p21	ENSG00000099260	palmdelphin
SQLE	Sqle	3.78196	chr8q24.1	ENSG00000104549	squalene epoxidase
RAG2	Rag2	3.78734	chr11p13	ENSG00000175097	recombination activating gene 2
AQP1	Aqp1	3.79434	chr7p14	ENSG00000106125	aquaporin 1 (Colton blood group)

	SMARCA2	Smarca2	3.79632	chr9p22.3	ENSG00000080503	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2
	PSAT1	Psat1	3.79808	chr9q21.2	ENSG00000135069	phosphoserine aminotransferase 1
	PTPRK	Ptprk	3.7984	chr6q22.2-23.1	ENSG00000152894	protein tyrosine phosphatase, receptor type, K
	TXN2	Txn2	3.80206	chr22q13.1	ENSG00000100348	thioredoxin 2
	TCF7	Tcf7	3.80427	chr5q31.1	ENSG00000081059	transcription factor 7 (T-cell specific, HMG-box)
	NCKAP1L	Nckap1l	3.80457	chr12q13.1	ENSG00000123338	NCK-associated protein 1-like
	XRCC1	Xrcc1	3.81405	chr19q13.2	ENSG00000073050	X-ray repair complementing defective repair in Chinese hamster cells 1
	NDUFA5	Ndufa5	3.81644	chr7q32	---	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 5, 13kDa
	CDKN2B	Cdkn2b	3.82074	chr9p21	ENSG00000147883	cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)
	TXNIP	Txnip	3.82367	chr1q21.1	ENSG00000117289	thioredoxin interacting protein
	PPP6C	Ppp6c	3.82451	chr9q33.3	ENSG00000119414	protein phosphatase 6, catalytic subunit
	PLXDC1	Plxdc1	3.82481	chr17q21.1	ENSG00000161381	plexin domain containing 1
	EHD4	Ehd4	3.82888	chr15q11.1	ENSG00000103966	EH-domain containing 4
	TCAP	Tcap	3.83883	chr17q12	ENSG00000173991	titin-cap (telethonin)
8 (49 genes)	SGK	Sgk	3.84073	chr6q23	ENSG00000118515	serum/glucocorticoid regulated kinase
	BLM	Blm	3.84131	chr15q26.1	ENSG00000197299	Bloom syndrome
	ZNF423	Zfp423	3.84204	chr16q12	---	Zinc finger protein 423
	NBL1	Nbl1	3.84563	chr1p36.13-p36.11	ENSG00000158747	neuroblastoma, suppression of tumorigenicity 1
	ITGA7	Itga7	3.85159	chr12q13	ENSG00000135424	integrin, alpha 7
	ENO3	Eno3	3.86052	chr17pter-p11	ENSG00000108515	enolase 3 (beta, muscle)

PACSIN2	Pacsin2	3.86178	chr22q13.2-13.33	ENSG00000100266	protein kinase C and casein kinase substrate in neurons 2
CENTD2	Centd2	3.86738	chr11q13.4	ENSG00000186635	centaurin, delta 2
MGAT1	Mgat1	3.87863	chr5q35	ENSG00000131446	mannosyl (alpha-1,3-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase
SDHB	Sdhb	3.88122	chr1p36.1-p35	ENSG00000117118	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
SRPX	Srpx	3.88385	chrXp21.1	ENSG00000101955	sushi-repeat-containing protein, X-linked
EGLN3	Egln3	3.88671	chr14q13.1	ENSG00000129521	egl nine homolog 3 (C. elegans)
POLR2L	Polr2l	3.89059	chr11p15	ENSG00000177700	polymerase (RNA) II (DNA directed) polypeptide L, 7.6kDa /// polymerase (RNA) II (DNA directed) polypeptide L, 7.6kDa
PMP22	Pmp22	3.89087	chr17p12-p11.2	ENSG00000109099	peripheral myelin protein 22
LTBP4	Ltbp4	3.89479	chr19q13.1-q13.2	ENSG00000090006	latent transforming growth factor beta binding protein 4
ITK	Itk	3.89795	chr5q31-q32	ENSG00000113263	IL2-inducible T-cell kinase
LSM3	Lsm3	3.89848	chr3p25.1	ENSG00000170860	LSM3 homolog, U6 small nuclear RNA associated (S. cerevisiae)
CD38	Cd38	3.90328	chr4p15	ENSG00000004468	CD38 molecule
CKAP4	Ckap4	3.90394	chr12q23.3	ENSG00000136026	cytoskeleton-associated protein 4
PHGDH	Phgdh	3.91476	chr1p12	ENSG00000092621	phosphoglycerate dehydrogenase
DLG7	Dlg7	3.9183	chr14q22.3	ENSG00000126787	discs, large homolog 7 (Drosophila)
EFHD1	Efhd1	3.93142	chr2q37.1	ENSG00000115468	EF-hand domain family, member D1
PCK2	Pck2	3.93339	chr14q12	ENSG00000100889	phosphoenolpyruvate carboxykinase 2 (mitochondrial)

HADHA	Hadha	3.93926	chr2p23	ENSG00000084754	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), alpha subunit
LOC387680	D6Wsu116e	3.94456			
SCARB2	Scarb2	3.95368	chr4q21.1	---	scavenger receptor class B, member 2
ACAT1	Acat1	3.95369	chr11q22.3-q23.1	ENSG00000075239	acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase)
TXN	Txn1	3.95531	chr9q31	---	Thioredoxin
VRK1	Vrk1	3.95618	chr14q32	ENSG00000100749	vaccinia related kinase 1
MX1	Mx2	3.96311	chr21q22.3	ENSG00000157601	myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse) /// myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)
OGDH	Ogdh	3.97278	chr7p14-p13	ENSG00000105953	oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)
SYPL1	Sypl	3.98562	chr7q22.2	ENSG00000008282	synaptophysin-like 1
SR140	2610101N10Rik	3.99061	chr3q23	---	U2-associated SR140 protein
APEX1	Apex1	3.99157	chr14q11.2-q12	ENSG00000100823	APEX nuclease (multifunctional DNA repair enzyme) 1
RABGAP1L	Rabgap1l	3.99435	chr1q24	ENSG00000152061	RAB GTPase activating protein 1-like
LAMB2	Lamb2	4.00081	chr3p21	ENSG00000172037	laminin, beta 2 (laminin S)
POLD2	Pold2	4.00496	chr7p13	---	Polymerase (DNA directed), delta 2, regulatory subunit 50kDa
RCN1	Rcn1	4.00714	chr11p13	---	reticulocalbin 1, EF-hand calcium binding domain
PDC	Pdc	4.00998	chr1q25.2	ENSG00000116703	phosducin

	VARS	Vars2	4.01065	chr6p21.3	ENSG00000204394 /// ENSG00000096171	valyl-tRNA synthetase
	GPC3	Gpc3	4.01352	chrXq26.1	ENSG00000147257	glypican 3
	CHPT1	Chpt1	4.01748	chr12q	ENSG00000111666	choline phosphotransferase 1
	GATA1	Gata1	4.01886	chrXp11.23	ENSG00000102145	GATA binding protein 1 (globin transcription factor 1)
	NOTCH3	Notch3	4.01929	chr19p13.2-p13.1	ENSG00000074181	Notch homolog 3 (Drosophila)
	NONO	Nono	4.01981	chrXq13.1	ENSG00000147140	non-POU domain containing, octamer-binding
	RTN1	Rtn1	4.02291	chr14q23.1	ENSG00000139970	reticulon 1
	ALOX5AP	Alox5ap	4.0268	chr13q12	ENSG00000132965	arachidonate 5-lipoxygenase-activating protein
	MEN1	Men1	4.02722	chr11q13	ENSG00000133895	multiple endocrine neoplasia I
	PRKAR1A	Prkar1a	4.03187	chr17q23-q24	ENSG00000108946	protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1)
9 (9 genes)	SPTA1	Spna1	4.03272	chr1q21	ENSG00000163554	spectrin, alpha, erythrocytic 1 (elliptocytosis 2)
	PDE1A	Pde1a	4.03709	chr2q32.1	---	phosphodiesterase 1A, calmodulin-dependent
	SLC4A1	Slc4a1	4.04481			
	PIGQ	Pigq	4.04696	chr16p13.3	ENSG00000007541	phosphatidylinositol glycan anchor biosynthesis, class Q
	MYOZ2	Myoz2	4.0483	chr4q26-q27	ENSG00000172399	myozenin 2
	SEC14L1	Sec14l1	4.05721	chr17q25.1-17q25.2	---	SEC14-like 1 (S. cerevisiae)
	MMD	Mmd	4.05742	chr17q	ENSG00000108960	monocyte to macrophage differentiation-associated
	ZNF160	Zfp26	4.05756	chr19q13.41	---	zinc finger protein 160
	CD3D	Cd3d	4.05909	chr11q23	ENSG00000167286	CD3d molecule, delta (CD3-TCR complex)
10 (38 genes)	GAP43	Gap43	4.06687	chr3q13.1-q13.2	---	growth associated protein 43
	ODC1	Odc1	4.06765	chr2p25	ENSG00000115758	ornithine decarboxylase 1

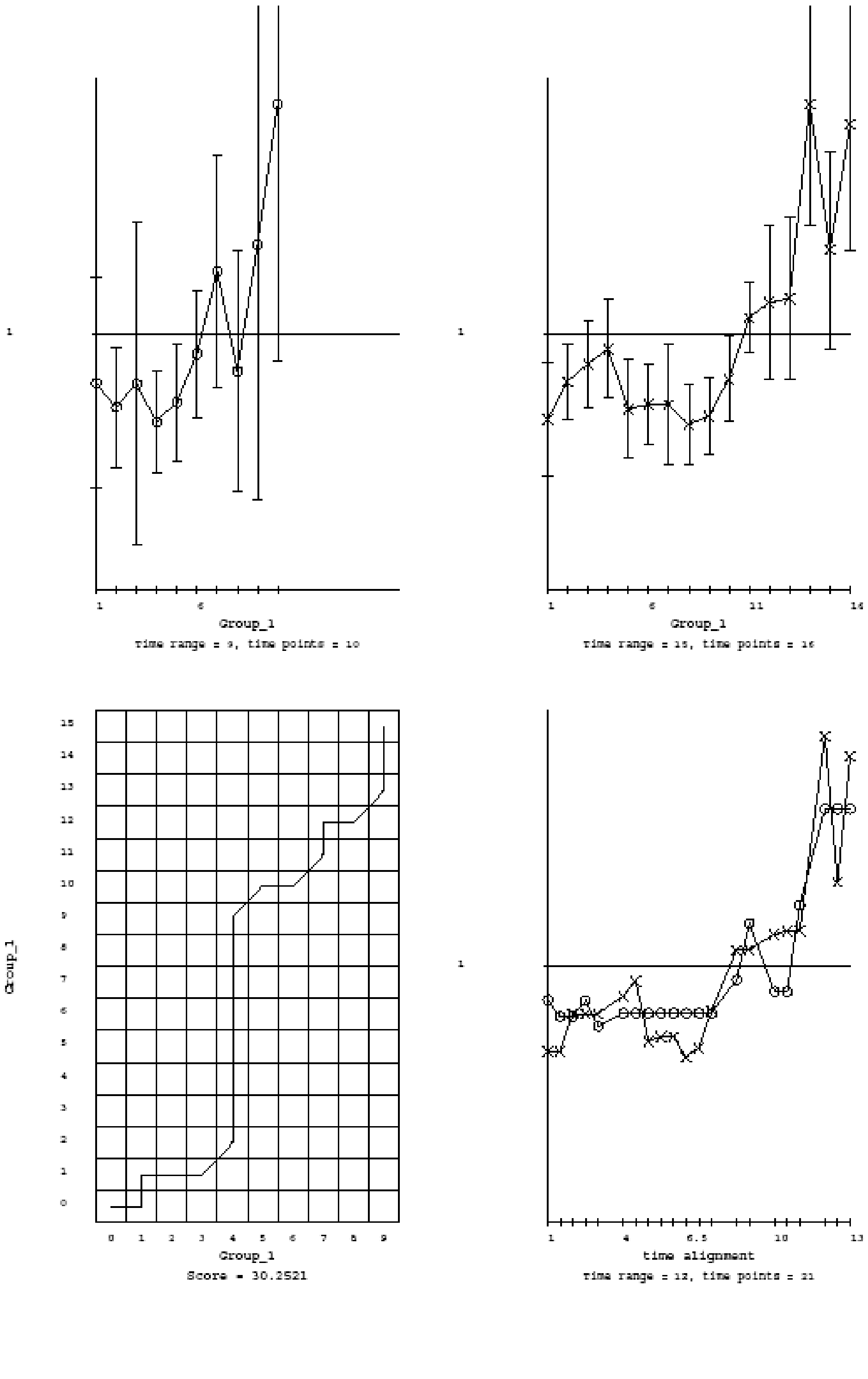
STAT3	Stat3	4.07063	chr17q21.31	ENSG00000168610	signal transducer and activator of transcription 3 (acute-phase response factor)
COL5A3	Col5a3	4.07122	chr19p13.2	ENSG00000080573	collagen, type V, alpha 3
KTN1	Ktn1	4.07466	chr14q22.1	ENSG00000126777	kinectin 1 (kinesin receptor)
PRKCD	Prkcd	4.0789	chr3p21.31	ENSG00000163932	protein kinase C, delta
PHLDA1	Phlda1	4.081	chr12q15	ENSG00000139289	pleckstrin homology-like domain, family A, member 1
UCHL1	Uchl1	4.08355	chr4p14	ENSG00000154277	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)
GALK1	Galk1	4.08428	chr17q24	ENSG00000108479	galactokinase 1
MDH1	Mdh1	4.0903	chr2p13.3	ENSG00000014641	malate dehydrogenase 1, NAD (soluble)
SUCLA2	Sucla2	4.09384	chr13q12.2-q13.3	ENSG00000136143	succinate-CoA ligase, ADP-forming, beta subunit
MLYCD	Mlycd	4.09803	chr16q24	ENSG00000103150	malonyl-CoA decarboxylase
SLC4A4	Slc4a4	4.10084	chr4q21	ENSG00000080493	solute carrier family 4, sodium bicarbonate cotransporter, member 4
RUFY2	Rufy2	4.10278	chr10q21.3	---	RUN and FYVE domain containing 2
HRC	Hrc	4.10447	chr19q13.3	ENSG00000130528	histidine rich calcium binding protein
ORC6L	Orc6l	4.10543	chr16q12	ENSG00000091651	origin recognition complex, subunit 6 like (yeast)
CAV1	Cav1	4.10718	chr7q31.1	ENSG00000105974	caveolin 1, caveolae protein, 22kDa
DNMT1	Dnmt1	4.10873	chr19p13.2	ENSG00000130816	DNA (cytosine-5-)-methyltransferase 1
NASP	Nasp	4.11372	chr1p34.1	ENSG00000132780	nuclear autoantigenic sperm protein (histone-binding)
RASSF3	Rassf3	4.11601	chr12q14.2	---	Ras association (RalGDS/AF-6) domain family 3
SNRPA1	Snrpa1	4.11936	chr15q26.3	ENSG00000131876	small nuclear ribonucleoprotein polypeptide A'

	SEC61B	Sec61b	4.12347	chr9q22.32-q31.3	ENSG00000106803	Sec61 beta subunit
	TBX5	Tbx5	4.12454	chr12q24.1	ENSG00000089225	T-box 5
	THBS1	Thbs1	4.12753	chr15q15	---	Thrombospondin 1
	NGFRAP1	Ngfrap1	4.13084	chrXq22.2	ENSG00000166681	nerve growth factor receptor (TNFRSF16) associated protein 1
	ARHGEF12	Arhgef12	4.1316	chr11q23.3	ENSG00000196914	Rho guanine nucleotide exchange factor (GEF) 12
	KLF9	Klf9	4.13334	chr9q13	ENSG00000119138	Kruppel-like factor 9
	WNT11	Wnt11	4.13515	chr11q13.5	ENSG00000085741	wingless-type MMTV integration site family, member 11
	RAP1A	Rap1a	4.13695	chr1p13.3	---	RAP1A, member of RAS oncogene family
	ILKAP	Ilkap	4.13996	chr2q37.3	ENSG00000132323	integrin-linked kinase-associated serine/threonine phosphatase 2C
	BSDC1	Bsdc1	4.13999	chr1p35.1	ENSG00000160058	BSD domain containing 1
	NPR3	Npr3	4.14203	chr5p14-p13	ENSG00000113389	natriuretic peptide receptor C/guanylate cyclase C (atrionatriuretic peptide receptor C)
	ABLIM1	Ablim1	4.14207	chr10q25	ENSG00000099204	actin binding LIM protein 1
	RPL14	Rpl14	4.14492	chr3p22-p21.2	---	ribosomal protein L14
	KLHL7	Klhl7	4.1459	chr7p15.3	ENSG00000122550	kelch-like 7 (Drosophila)
	CITED1	Cited1	4.14591	chrXq13.1	ENSG00000125931	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 1
	SP100	Sp100	4.14747	chr2q37.1	---	SP100 nuclear antigen
	CDC25A	Cdc25a	4.15595	chr3p21	ENSG00000164045	cell division cycle 25 homolog A (S. cerevisiae)
11 (9 genes)	CDIPT	Cdipt	4.15666	chr16p11.2	ENSG00000103502	CDP-diacylglycerol--inositol 3-phosphatidyltransferase (phosphatidylinositol synthase)

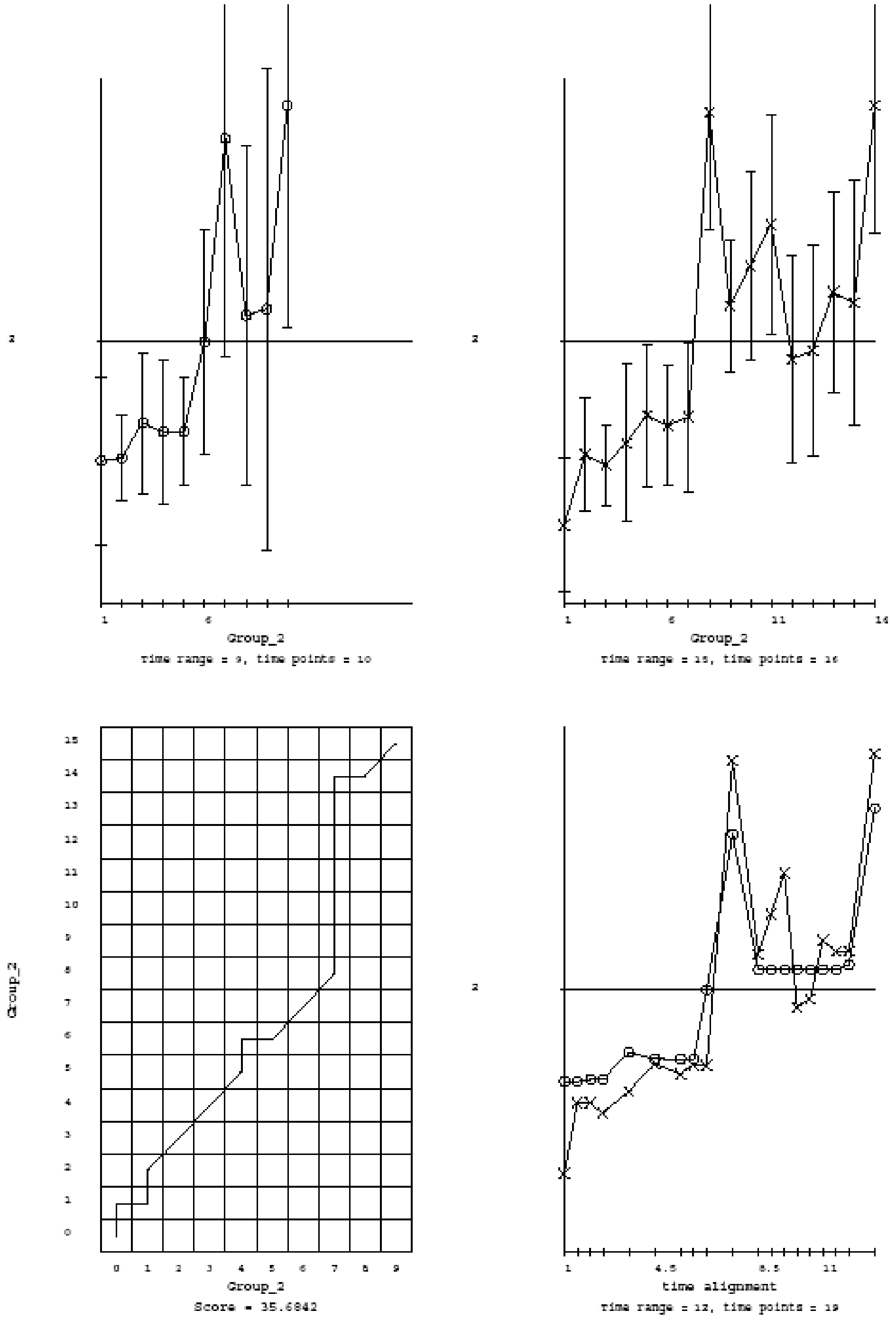
	PRKD2	Prkd2	4.15827	chr19q13.3	ENSG00000105287	protein kinase D2
	TALDO1	Taldo1	4.15958	chr11p15.5-p15.4	ENSG00000177156	transaldolase 1
	GENX-3414	D5Ertd593e	4.16432	chr4q24-q25	---	genethonin 1
	LMNB2	Lmnb2	4.16501	chr19p13.3	ENSG00000176619	lamin B2
	SYN1	Syn1	4.16552	chrXp11.23	ENSG00000008056	synapsin I
	ESRRG	Esrrg	4.16678	chr1q41	ENSG00000196482	estrogen-related receptor gamma
	CDC6	Cdc6	4.16971	chr17q21.3	ENSG00000094804	cell division cycle 6 homolog (S. cerevisiae)
	TNFRSF4	Tnfrsf4	4.17453	chr1p36	ENSG00000186827	tumor necrosis factor receptor superfamily, member 4
12 (7 genes)	PDE2A	Pde2a	4.1751	chr11q13.4	ENSG00000186642	phosphodiesterase 2A, cGMP-stimulated
	PPOX	Ppox	4.17989	chr1q22	ENSG00000143224	protoporphyrinogen oxidase
	HSPA5	Hspa5	4.19148	chr9q33-q34.1	ENSG00000044574	heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)
	CYFIP2	Cyfip2	4.19543	chr5q33.3	ENSG00000055163	cytoplasmic FMR1 interacting protein 2 /// cytoplasmic FMR1 interacting protein 2
	HIST1H2BG	Hist1h2bm	4.19658	chr6p21.3	ENSG00000187990	histone cluster 1, H2bg
	AP1B1	Ap1b1	4.19898	chr22q12|22q12.2	ENSG00000100280	adaptor-related protein complex 1, beta 1 subunit
	HNRPM	Hnrpm	4.19951	chr19p13.3-p13.2	---	heterogeneous nuclear ribonucleoprotein M

Appendix B

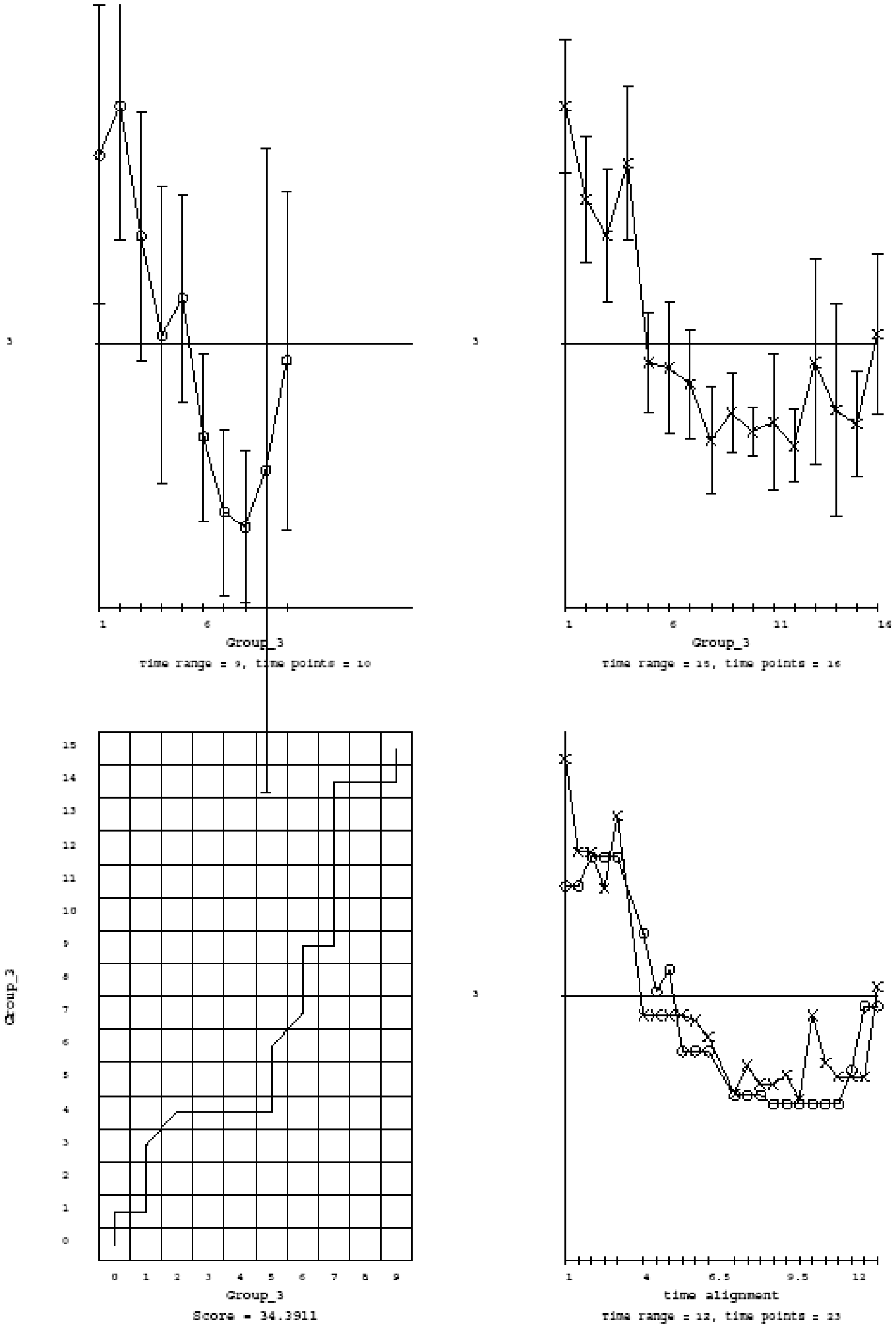
Cluster 1



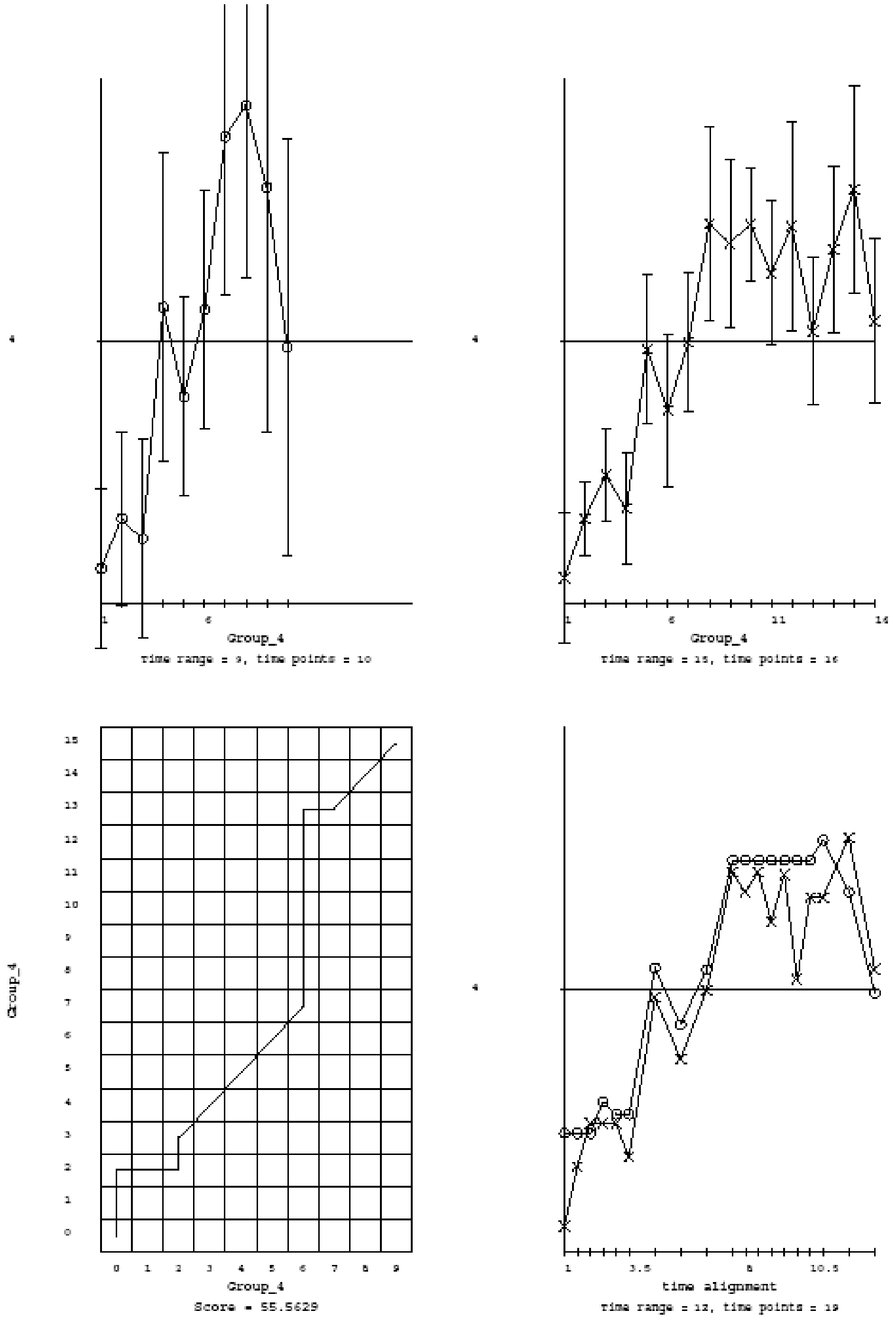
Cluster 2


2
1
6
Group_2
Time range = 9, time points = 10
2
1
6
11
16
Group_2
Time range = 15, time points = 16
Group_2
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
Group_2
Score = 35.6842
2
1
4.5
8.5
11
time alignment
Time range = 12, time points = 19

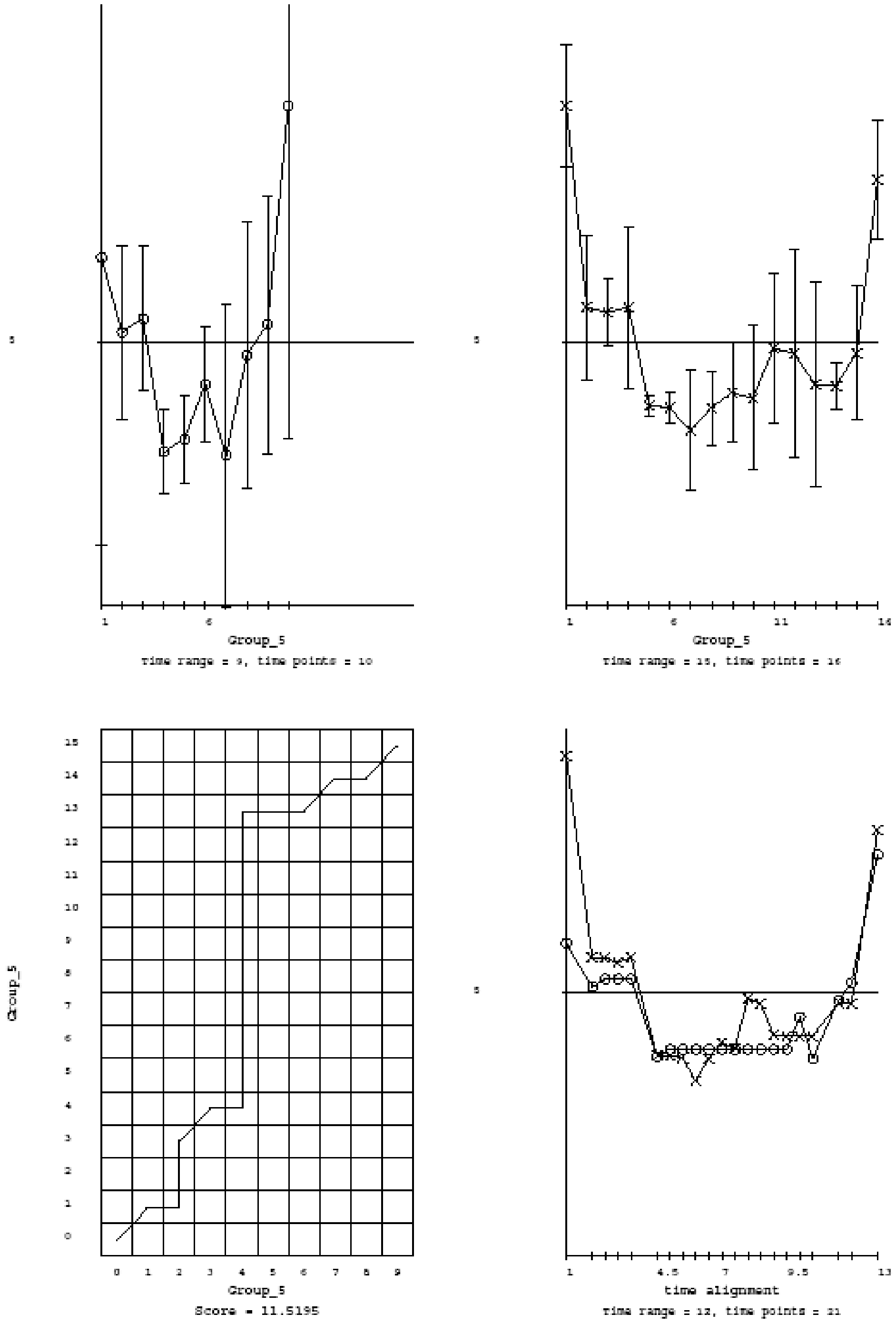
Cluster 3


3
1
6
Group_3
Time range = 9, time points = 10
3
1
6
11
16
Group_3
Time range = 15, time points = 16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Group_3
0
1
2
3
4
5
6
7
8
9
Group_3
Score = 34.3911
3
1
4
6.5
9.5
12
time alignment
Time range = 12, time points = 23

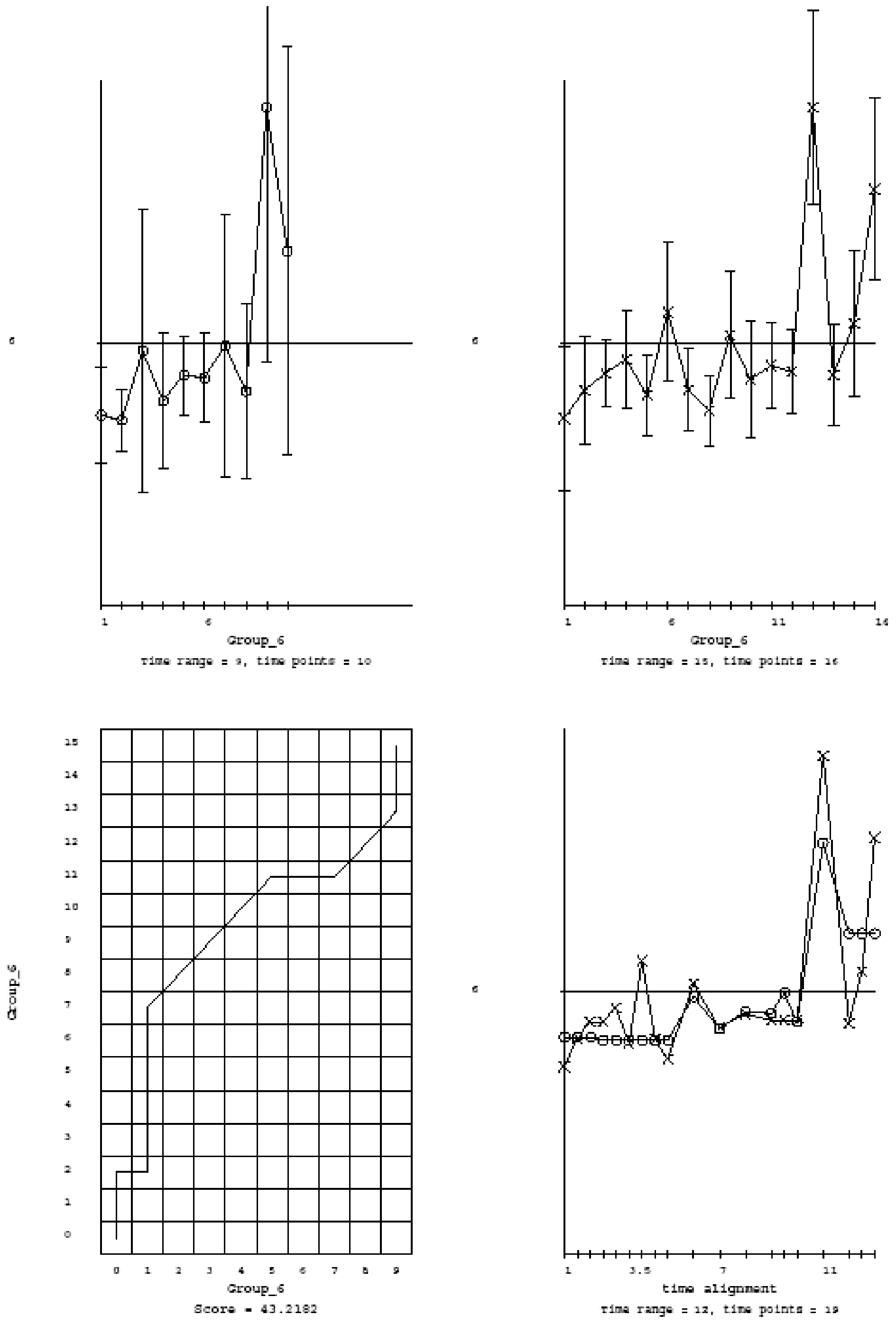
Cluster 4


4
1
6
Group_4
time range = 9, time points = 10
4
1
6
11
16
Group_4
time range = 15, time points = 16
Group_4
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
Group_4
Score = 55.5629
4
1
3.5
8
10.5
time alignment
time range = 12, time points = 19

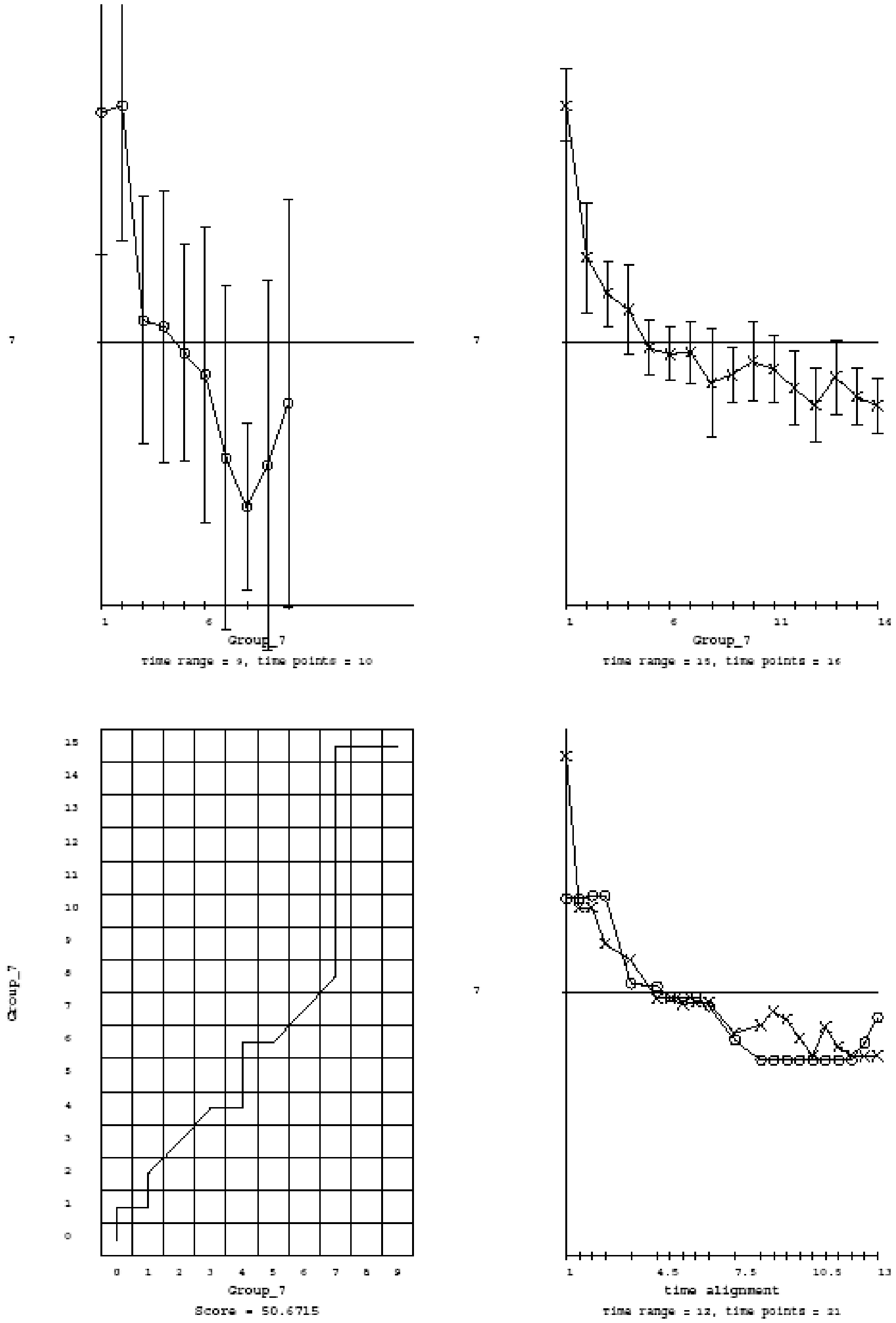
Cluster 5


5
1
6
Group_5
Time range = 9, time points = 10
5
1
6
11
16
Group_5
Time range = 15, time points = 16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Group_5
0
1
2
3
4
5
6
7
8
9
Group_5
Score = 11.5195
5
1
4.5
7
9.5
13
time alignment
Time range = 12, time points = 21

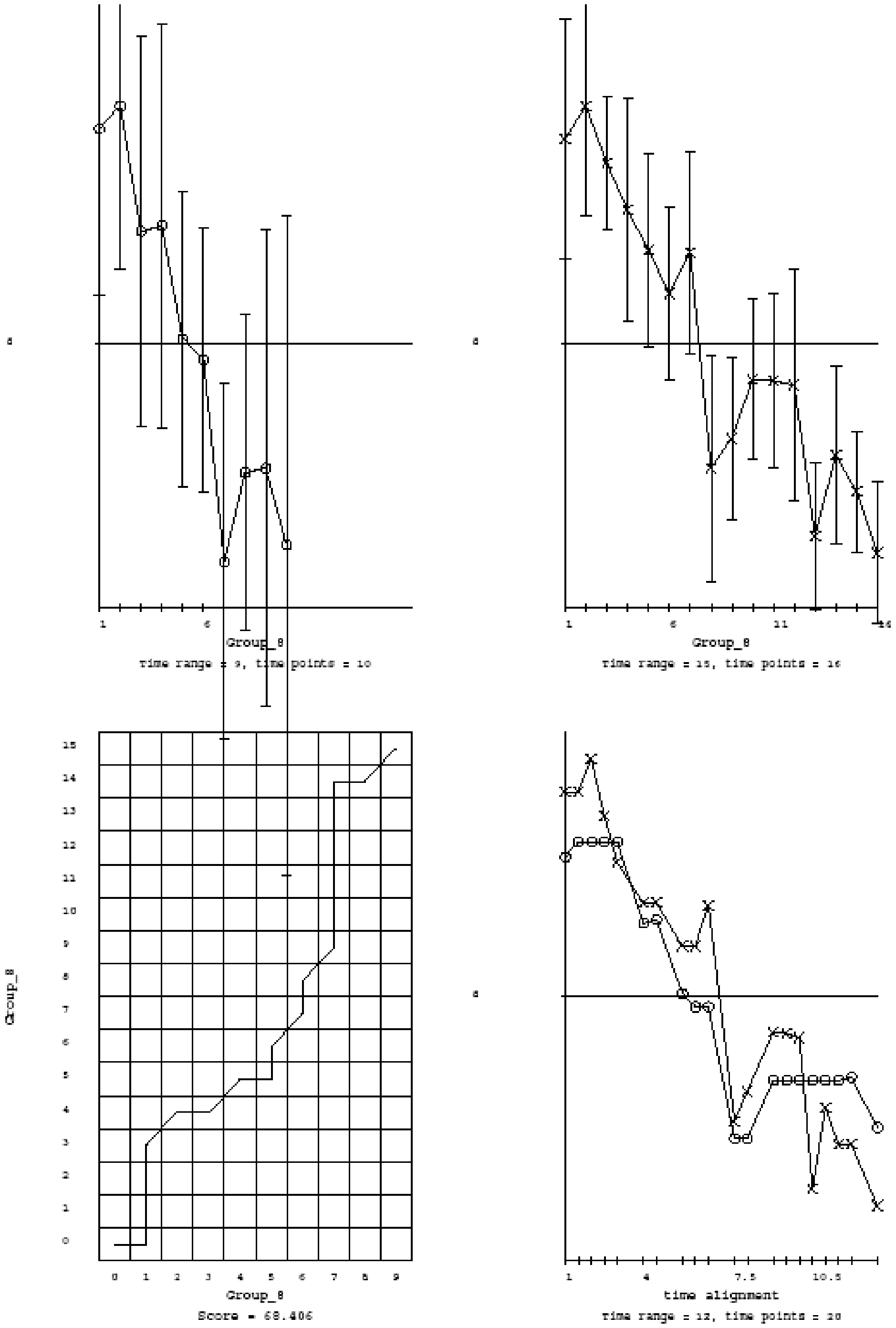
Cluster 6


6
1
6
Group_6
Time range = 9, time points = 10
6
1
6
11
16
Group_6
Time range = 15, time points = 16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Group_6
0
1
2
3
4
5
6
7
8
9
Group_6
Score = 43.2182
6
1
3.5
7
11
time alignment
Time range = 12, time points = 19

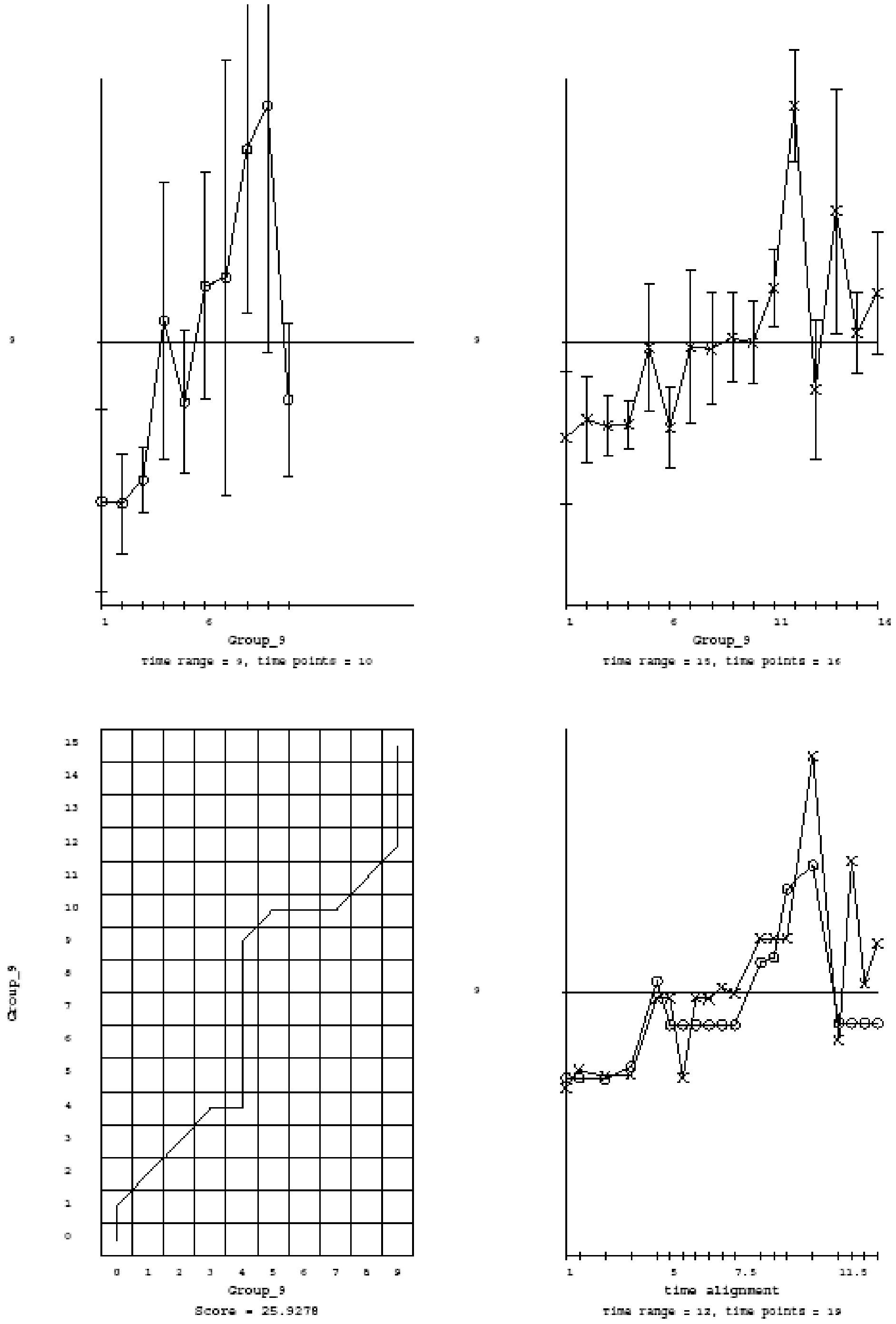
Cluster 7


7
1
6
Group_7
time range = 9, time points = 10
7
1
6
11
16
Group_7
time range = 15, time points = 16
Group_7
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9
Group_7
Score = 50.6715
7
1
4.5
7.5
10.5
13
time alignment
time range = 12, time points = 21

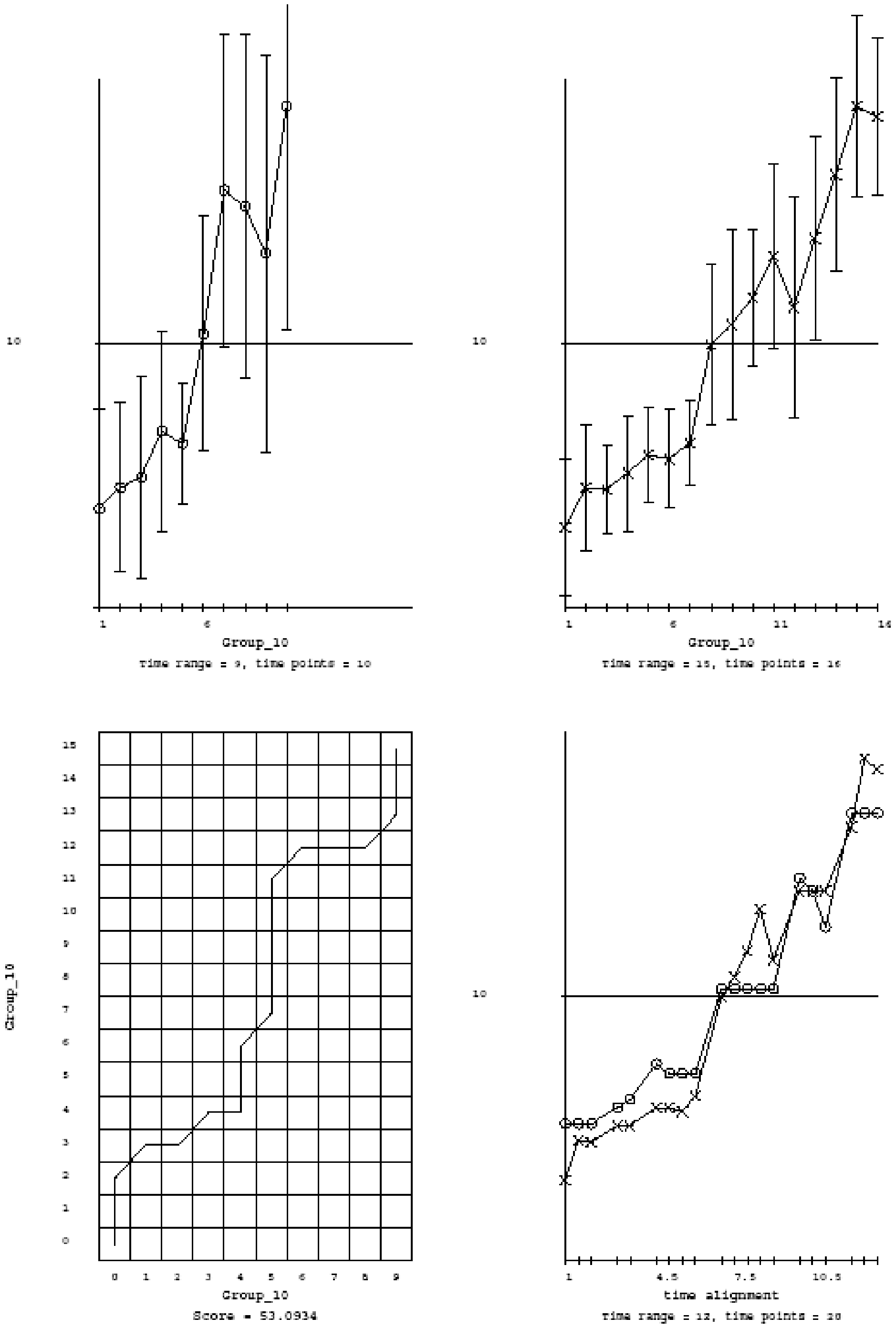
Cluster 8


0
1
6
Group_8
Time range = 9, time points = 10
0
1
6
11
16
Group_8
Time range = 15, time points = 16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Group_8
0
1
2
3
4
5
6
7
8
9
Group_8
Score = 68.406
0
1
4
7.5
10.5
time alignment
Time range = 12, time points = 20

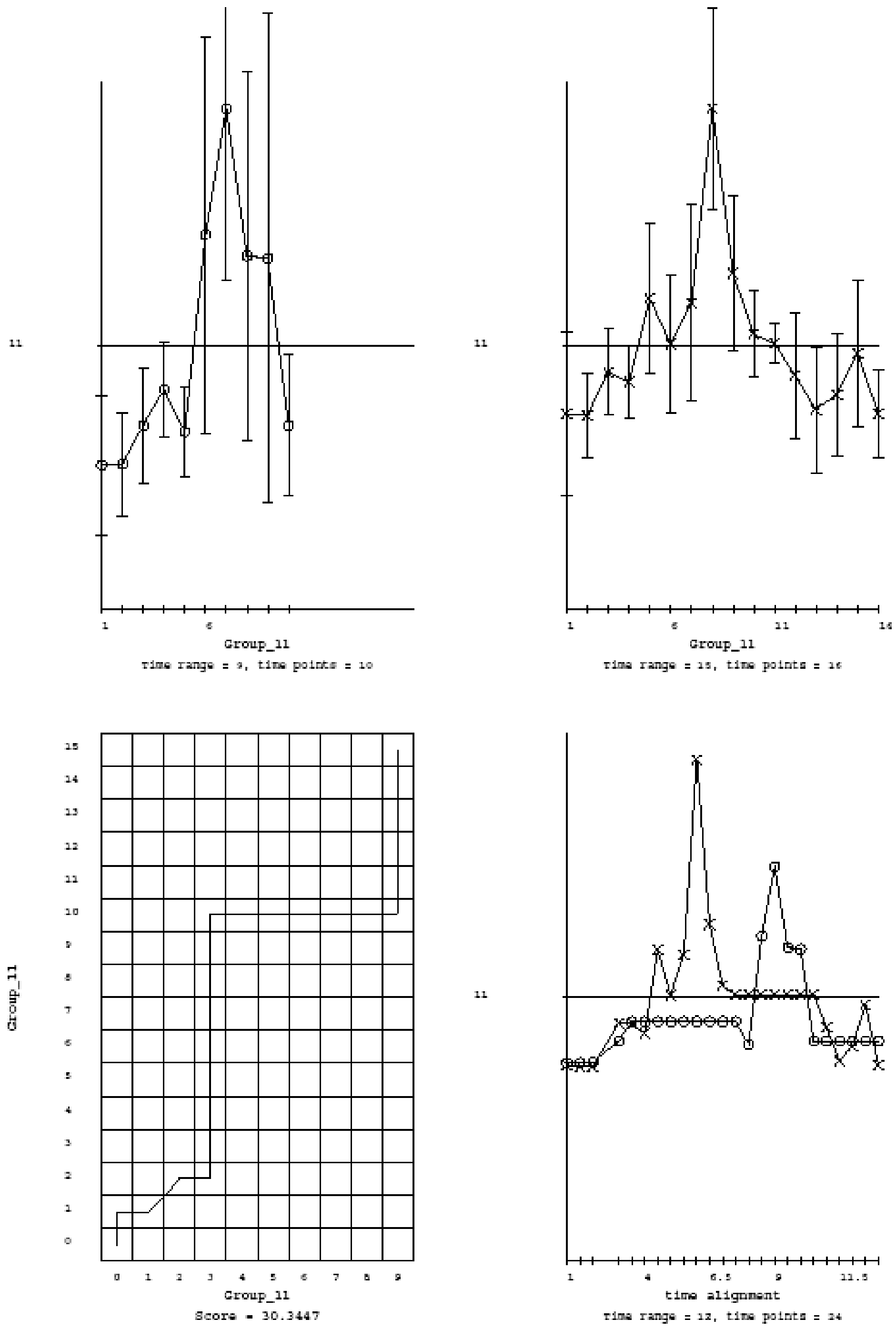
Cluster 9


9
1
6
Group_9
Time range = 9, time points = 10
9
1
6
11
16
Group_9
Time range = 15, time points = 16
Group_9
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
Group_9
Score = 25.9278
9
1
5
7.5
11.5
time alignment
time range = 12, time points = 19

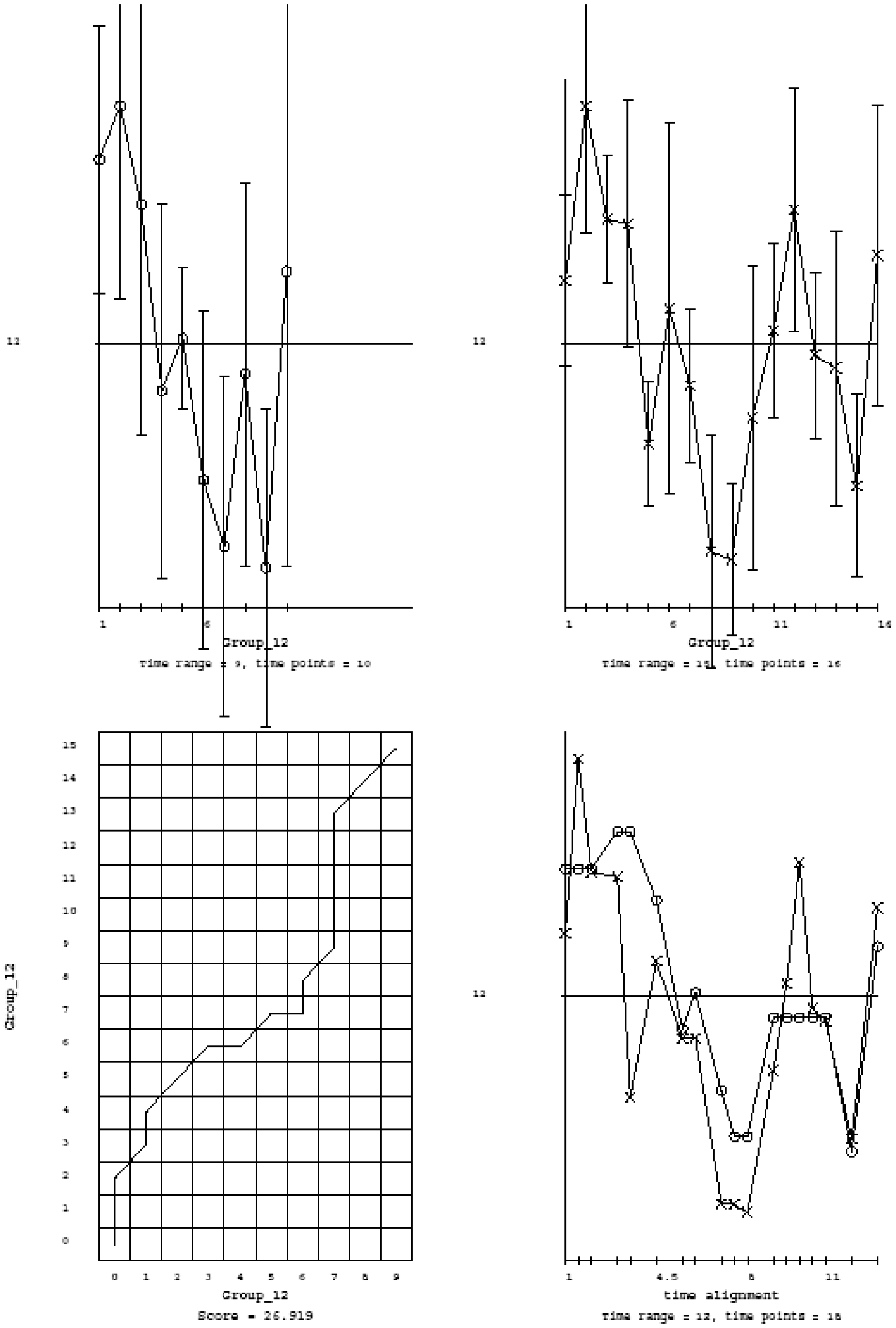
Cluster 10


10
1
6
Group_10
Time range = 9, time points = 10
10
1
6
11
16
Group_10
Time range = 15, time points = 16
Group_10
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
Group_10
Score = 53.0934
10
1
4.5
7.5
10.5
time alignment
Time range = 12, time points = 20

Cluster 11


11
1
6
Group_11
Time range = 9, time points = 10
11
1
6
11
16
Group_11
Time range = 15, time points = 16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Group_11
0
1
2
3
4
5
6
7
8
9
Group_11
Score = 30.3447
11
1
4
6.5
9
11.5
time alignment
Time range = 12, time points = 24

Cluster 12


12
1
6
Group_12
Time range = 9, time points = 10
12
1
6
11
16
Group_12
Time range = 15, time points = 16
Group_12
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9
Group_12
Score = 26.919
12
1
4.5
8
11
time alignment
Time range = 12, time points = 18