Research

Semi-supervised gene shaving method for predicting low variation biological pathways from genome-wide data
Dongxiao Zhu[1,2]

Address: [1]Department of Computer Science, University of New Orleans, New Orleans, LA 70148, USA and [2]Research Institute for Children, Children's Hospital, New Orleans, LA 70118, USA

Email: Dongxiao Zhu - dzhu@cs.uno.edu

This article is available from: http://www.biomedcentral.com/1471-2105/10/S1/S54

Abstract

Background: The gene shaving algorithm and many other clustering algorithms identify gene clusters showing high variation across samples. However, gene expression in many signaling pathways show only modest and concordant changes that fail to be identified by these methods. The increasingly available signaling pathway prior knowledge provide new opportunity to solve this problem.

Results: We propose an innovative semi-supervised gene clustering algorithm, where the original gene shaving algorithm was extended and generalized so that prior knowledge of signaling pathways can be incorporated. Different from other methods, our method identifies gene clusters showing concerted and modest expression variation as well as strong expression correlation. Using available pathway gene sets as prior knowledge, whether complete or incomplete, our algorithm is capable of forming tightly regulated gene clusters showing modest variation across samples. We demonstrate the advantages of our algorithm over the original gene shaving algorithm using two microarray data sets. The stability of the gene clusters was accessed using a jackknife approach.

Conclusion: Our algorithm represents one of the first clustering algorithms that is particularly designed to identify signaling pathways of low and concordant gene expression variation. The discriminating power is achieved by manufacturing a principal component enriched by signaling pathways.

Background
Gene clustering that assigns group membership(s) to each gene is a widespread pattern extraction technique. Genes sharing the same membership are often hypothesized to be regulated by the same defined or undefined genomic influence, such as cellular pathway. Model-free clustering techniques such as K-means and hierarchical clustering [1-3] are widely used. One limitation of these approaches, as pointed out by many researchers, e.g. [4], is that each gene can only belong to a single cluster. These types of gene clustering algorithms are thus called mutually exclusive clustering. In the context of cellular pathways, they assume that one gene can only be regulated by one pathway at a time, which apparently, is not the case. Model-

based clustering or soft clustering [5-8] provides mechanisms to relax this stringent assumption by introducing "probabilistic" or "fuzzy" memberships. However, these "soft" memberships do not biologically account for the fact that one gene is often simultaneously regulated by multiple genomic influences.

Singular value decomposition (SVD) [9-11] has shown great promise towards deconvolving channels of genomic influence. Assuming rows of data matrix correspond to genes and columns correspond to physiological/genetic conditions under which the gene expression abundance was interrogated using gene chips, the SVD factors the data matrix into three matrices. The first matrix, which contains most of information, is called a gene coefficient matrix where each column (principal component, PC) defines a preliminary gene cluster that might be regulated by a specific genomic influence. We will describe more details of SVD in the method section. SVD has been repeatedly shown to be able to deconvolve the observed gene expression signal into a composite of multiple overlapping genomic influences, many of them correspond to signaling pathways [9,11].

Thus SVD provides a methodology base for non-mutually exclusive clustering. The gene clusters generated by SVD are often preliminary due to the fact that many non-relevant genes might contaminate the PC's that define gene clusters. Hastie et al [4] proposed removing non-relevant genes in an iterative fashion, in which the least correlated genes with the leading PC is treated as non-relevant. The gene shaving algorithm quickly became an important tool in the pattern discovery arsenal. It iteratively searches for clusters of genes showing high variation across the samples, and correlation across the genes [8]. The former is achieved by working with the leading PC and the latter is achieved by iteratively discarding non-relevant genes to the cluster. There are other types of non-mutually exclusive clustering methods as well, such as plaid model [12].

The underlying assumption of the gene shaving algorithm is that the leading PC accounting for the largest portion of variation is always of exclusive interest to the investigator [4,13]. Consequently the algorithm iteratively refines the first gene cluster defined by the first PC by shaving off a proportion of genes that are least correlated with the leading PC. The second gene cluster is formed by performing the same procedure on the orthogonal data, resulting from the residuals of regression, and so on. However, the underlying assumption that the whole algorithm is based on is not always true for every single case. In fact, gene expression in many signaling pathways show modest but concordant changes. The gene shaving algorithm would most likely to fail in these cases by working exclusively with the leading PC.

Gene set based methods, such as Gene Set Enrichment Analysis (GSEA) were designed to overcome this limitation. Since it's first introduction in 2003 [14], it has been widely applied to interpret genome-wide expression profiles [15,16]. However, the approach only ranks pre-compiled gene sets according to the relevancy to the data and does not predict any new genes in the gene sets. Therefore, it strictly depends on the availability and validity of *a priori* defined gene sets. In reality a gene set is not always available in a complete and accurate format. What is typically available is partial pathway learned from empirical experimental studies.

We seek a seamless combination of the strengths of the two methodological frameworks. We manufacture a PC that is most enriched by prior knowledge (signaling pathway of interest). Performing the analysis iteratively we will be able to identify the gene cluster showing modest but concordant changes. In many cases, we are further interested in finding genes that are concordantly up or down-regulated by genomic influences. Therefore, it might be beneficial to turn our attention not only to the PC that the prior knowledge is most enriched, but also to the positive PC and the negative PC separatively. The hypothesis can be substantiated by previous works that positive and negative PC's can be enriched by completely different biological functions, e.g. [11].

In our work, we eliminate non-relevant genes iteratively following and improving the procedure used in the gene shaving algorithm [4]. In each iteration, a weighted average expression profile was calculated and used as the seed profile to rank genes. With the heuristic removal of non-
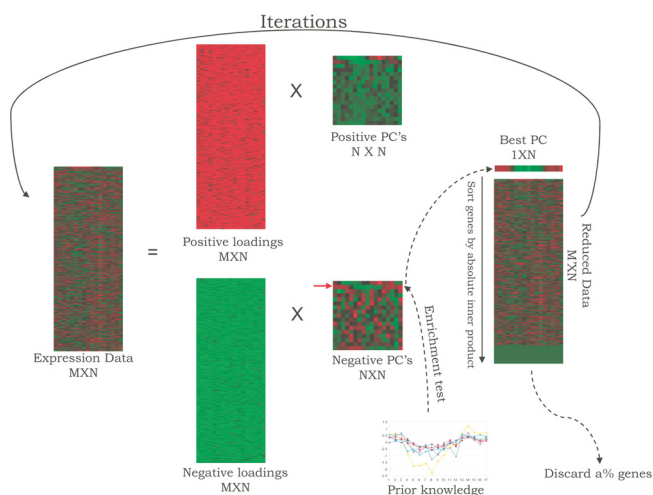

Figure 1

The schematic diagram of the proposed algorithm. "Enrichment test" means to determine the PC(s) that are most enriched by a prior knowledge gene set. α% is set to 10% following Hastie et al [4].

relevant genes at the beginning of the iterations, and some relevant genes by the end, the enrichment of prior knowledge has seen a sharp increase, followed by a gradual decrease. We then propose a trace-back step to retrieve the gene cluster in which enrichment of prior knowledge is maximized (Figure 1).

Results

We aim to demonstrate that the proposed algorithm is capable of identifying tightly regulated gene sets showing modest and concerted variation using incomplete prior knowledge and real-world microarray data set. Ground truth, which indicates a "complete" gene set used as precondition for applying GSEA algorithm [14,16], is desirable to demonstrate the claimed advantages of our algorithm. It is often not available. Therefore, we use four "high-amplitude" and four "low-amplitude" gene sets identified in [17] as ground truth to evaluate the ability of our algorithm to recover them using subsets of a variety of lengths. The high and low amplitude genes used in this example are well-studied genes in the cell cycle, and many of them are co-regulated by a number of signaling pathways [17,18]. We then use incomplete prior knowledge supplied by our collaborator and apply our algorithm to predict new WNT and NOTCH pathway genes in the somitogenesis process.

Recovering low and high amplitude gene sets using incomplete prior knowledge

As a proof of concept, we first analyzed a cell cycle data set originally reported in [17]. The data set consists of whole yeast genome expression profiles interrogated over two full cell cycles (20 evenly spaced time points) synchronized by elutriation. We considered the same 308 genes as in the paper derived using Fourier transform. In each of the four gene sets, genes were further classified into high-amplitude and low-amplitude groups according to magnitude of variation. The processed data are available from the authors' website at [19].

We treated the high-amplitude genes and low-amplitude genes in each gene set as "complete", as assumed in classical GSEA analysis. We sampled subsets of increasing sizes from 5 to complete (e.g. 40) with a step size of 5. In each step experiment, we generated 500 subsets of the same size (with replicates), and for each subset we applied our algorithm to demonstrate its ability to recover the full gene set using the hypergeometric test explained in method section. The *P*-values of the tests were used as a measure for such an ability. For visualization convenience, the *P*-values were negatively log-transformed and higher value corresponds to better recovery of the complete gene set.

The high-amplitude and low-amplitude complete gene sets were plotted in Figure 2a (upper panel of Figure 2). In both Fig 2b (lower left panel of Figure 2) and Fig 2c (low right panel of Figure 2), the ability of recovering the complete gene set (ground truth) was plotted against the increasing subset size respectively. The observed monotonic increase indicates that the larger the subsets (prior knowledge) are, the more capable of recovering the complete gene set. It is worth mentioning that Figure 2b demonstrates the capability of our algorithm to recover low-amplitude gene set, and Figure 2c demonstrates the capability of the gene shaving algorithm [4] to recover high-amplitude gene set.

Our algorithm can be viewed as an generalization of the gene shaving algorithm. Gene shaving algorithm exclusively works with the leading PC. Therefore, it is only capable of identifying high-amplitude signaling pathways. Our algorithm adaptively works with the PC that is most enriched by prior knowledge. Therefore, it is capable of identifying either high-amplitude or low-amplitude signaling pathways wherever prior knowledge is available. Comparing Figure 2b to Figure 2c more closely, it is evident that our algorithm recovers low-amplitude gene sets even better than gene shaving algorithm recovers high-amplitude ones. This is demonstrated by uniformly larger mean values and overall smaller variance on the vertical axis. The results of analyzing other complete gene sets of appropriate size lead to the same conclusion (see additional file 1). The proof-of-concept analysis provided compelling evidence that our algorithm is particularly suitable for identifying sets of tightly regulated genes with modest variation.

Predicting WNT and NOTCH pathway genes using prior knowledge

Microarray data and prior knowledge

We then proceed to re-analyze microarray data originally reported in Dequeant et al [20] to predict genes in WNT and NOTCH pathways. In this experiment, the genome-wide gene expression was interrogated over 17 developmental stages using Affymetrix GeneChip 430A. Using the Lomb-Scargle periodogram [21] the top 687 genes were used for gene clustering so that all prior knowledge genes are included. Microarray data are available at ArrayExpress at [22].

Prior knowledge corresponds to a list of experimentally validated cyclic genes regulated by the segmentation clock, a molecular oscillator acting during somitogenesis [20]. The segmentation clock is a set of periodic processes linked to the formation of the vertebrate embryo segments (somites) that give rise to the segments in the adult body plan of a vertebrate animal. Malfunction of cyclic genes are the direct cause of many developmental dis-
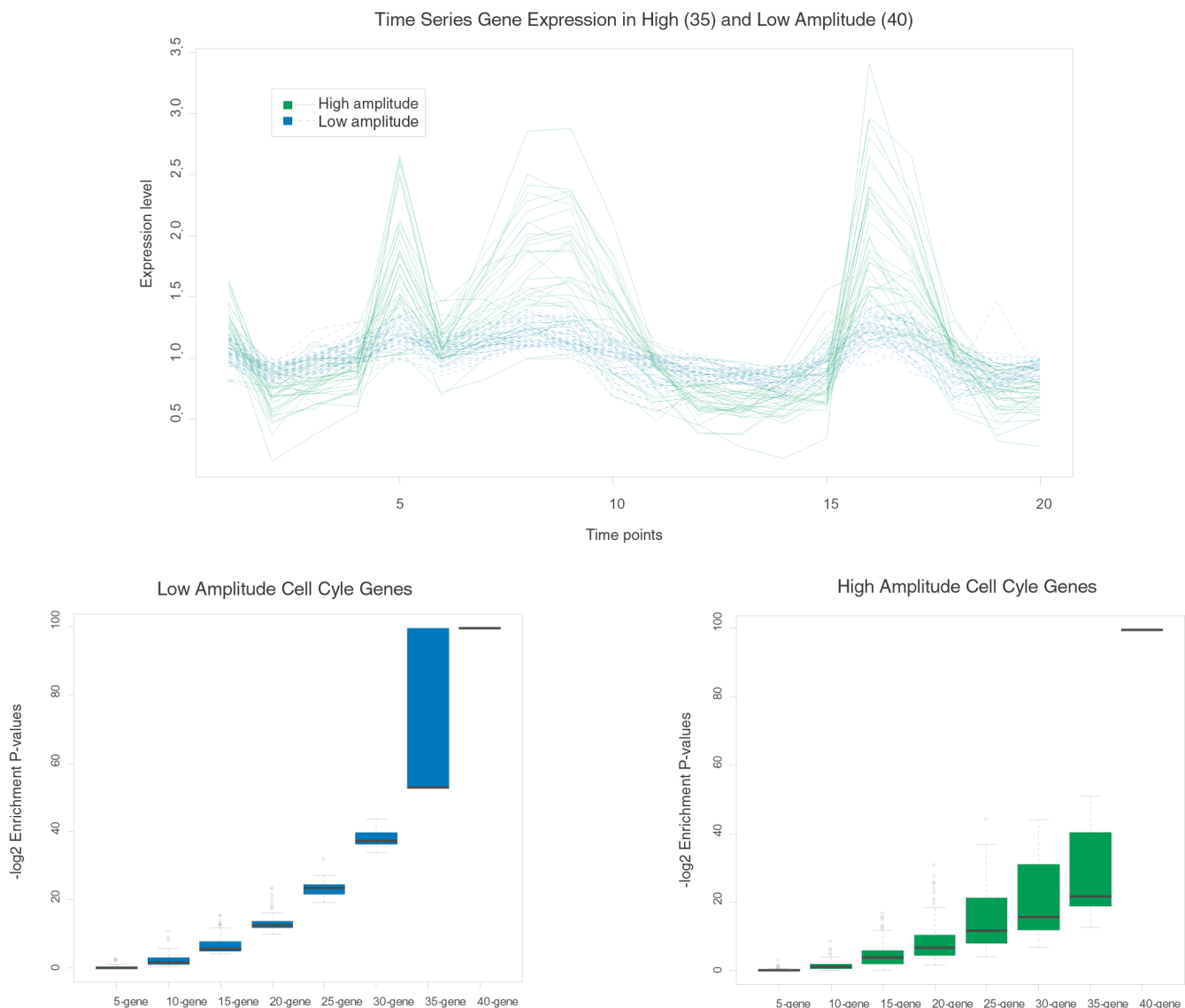
Figure 2
Demonstration of the claimed advantages of our algorithm using the "ground truth" reported in [17]. (a) Plots of expression profiles of high-amplitude and low-amplitude gene sets. (b) Evaluating the capability of our algorithm to recover a complete low-amplitude gene set. The gene shaving shaving algorithm [4] fails in this case because it exclusively works with the leading PC. X-axis represents the increasing sizes of the subsets, and Y-axis represents the *-log2P* of the enrichment, indicating increased capacity of recovering a complete gene set. (c) Evaluating the capability of gene shaving algorithm [4] to recover a complete high-amplitude gene set.

eases, such as Noonan syndrome and truncated tail [20]. Therefore, predicted cyclic genes are potential human disease genes. In particular, we have incomplete sets of 11 genes in the WNT pathway, and 9 genes in the NOTCH pathway as our prior knowledge. Our objective is to predict more WNT and NOTCH genes using prior knowledge, microarray data and our proposed algorithm.

Finding the most enriched PC using prior knowledge
In each iteration of our algorithm, we search for the PC that is most enriched by known WNT and NOTCH genes. We filtered the gene coefficients in each PC using the cutoff and tested enrichment of known pathway genes using the hypergeometric test (see method section). Figure 3 shows what happened in the first iteration where all 11 known WNT genes and all 9 known NOTCH genes are included in the second PC (enrichment level is $E - 06$).

After separating positive and negative PC's, in Figure 4, all known WNT genes are included in the second negative PC and all known NOTCH genes are included in the second positive PC (enrichment level is $E - 10$). The marked increase of P-value reveals that separating positive PC from negative PC is a key to better enrichment of prior knowledge. The fact that prior knowledge is mostly enriched in PC's other than the leading one indicates that the gene expression in the NOTCH and WNT pathways show only modest and concordant changes. The enrichment of prior knowledge in the gene cluster could be further improved as our algorithm iterates. In the next section, we present results of generating the "best" WNT and NOTCH clusters in which enrichment of prior knowledge is optimized.

Comparing our semi-supervised algorithms with the gene shaving algorithm
We aim to show that our semi-supervised algorithm is uniquely able to identify low variation signaling pathway genes but not the gene shaving algorithm. For predicting WNT cluster, our algorithm terminates after 18 iterations, and for predicting NOTCH cluster, it terminates after 20 iterations. We then traced back to retrieve the optimized clusters. Both WNT and NOTCH clusters were retrieved at the 9th iteration that prior knowledge is most enriched, and were smallest clusters containing all prior knowledge genes (Figure 5c). From Figure 5a, the original gene shaving algorithm [4] apparently failed in this case demonstrated by no enrichment of prior knowledge at all. The reason is, as discussed before, that WNT and NOTCH pathway genes are concordantly regulated in modest magnitude while gene shaving algorithm only works with the leading PC. Figures 5b and 5c present the prior knowledge enrichment achieved by two variants of our semi-supervised algorithm: with or without separating positive PC's from negative PC's. It is evident that splitting PC's gives rise to better clustering performance.

The left panel of Figure 6 plots gene expression profiles of the predicted NOTCH cluster, and right panel displays the annotation of those genes. Genes in the shaded areas are from our prior knowledge [20], and genes that are pointed by red arrows indicate the genes are experimentally validated to be positive, and genes pointed by blue arrows indicate the genes are potentially relevant through literature search. Note that the two pathways are far less from well understood, and therefore, many predicted genes, although not currently supported by experimental evidence, are likely to be validated later.

To make our prediction useful for improving current understanding of the mechanisms of WNT and NOTCH pathways in somatogenesis, we performed analysis to infer what kinds of biological functions (defined by Gene Ontology, GO) are most enriched in the pathways, and what kind of transcription factors (inferred through ChIP-chip experiments) are most likely to be involved in regulating the two pathways. Table 1 presents the results of abovementioned enrichment analysis. The analysis was done through the web-server of the Segal lab: [23]. In table 1, results appear to be meaningful since many significantly enriched GO terms (column 3) are related to embryonic development, and both enriched transcription factors (column 4): MyoG and MyoD are closely related to cell differentiation [24,25]. In particular, Myod and Myog have distinct regulatory roles at a similar set of target genes. The role of Myog in mediating terminal differentiation is partially to enhance expression of a subset of genes previously turned on by Myod [25].

Stability of clusters against perturbation of prior knowledge
Our approach predicts new pathway genes based on the available prior knowledge, therefore, it is critical to investigate the sensitivity of our prediction to a modest perturbation of prior knowledge. Since in this data set we don't know such ground truth as we did in the cell cycle data analysis, we performed sensitivity analysis using leave-one-out and leave-two-out jackknife approaches, see method section for technical details. Narrower Jackknife confidence interval of the enrichment indicates better stability of our enrichment estimation against perturbation of prior knowledge. In Figure 7a where the leave-one-out approach was applied, the estimation of enrichment is perfectly stable (zero variance) and increases until the ninth iteration. Recall that we traced back and retrieved the "best" NOTCH gene cluster right in the ninth iteration. This translates into the fact that our cluster analysis is very robust against moderate perturbation of prior knowledge. In Figure 7b where the leave-two-out approach was used follows a similar trend but with better stability (a narrower confidence interval). This is due to the fact that there are a larger number of Jackknife samples available in leave-two-out approach.

Discussion
With exception of a few recent works [26-28], most clustering algorithms these days are non-supervised in the sense that prior knowledge is not properly utilized to guide the learning process. Instead prior knowledge is often used in the post-learning phase in that researchers predict functions of unknown genes based on genes of known functions lying in the same cluster. The traditional gene shaving method focuses on the leading PC that accounts for most of variation in the data. On one hand, it is useful in discovering high variation pathway genes [4,29], on the other hand, it tends to overlook essential pathway genes that have modest expression variation. We hypothesized that highly concerted expression behavior of these genes, albeit modest in variation, may help shape
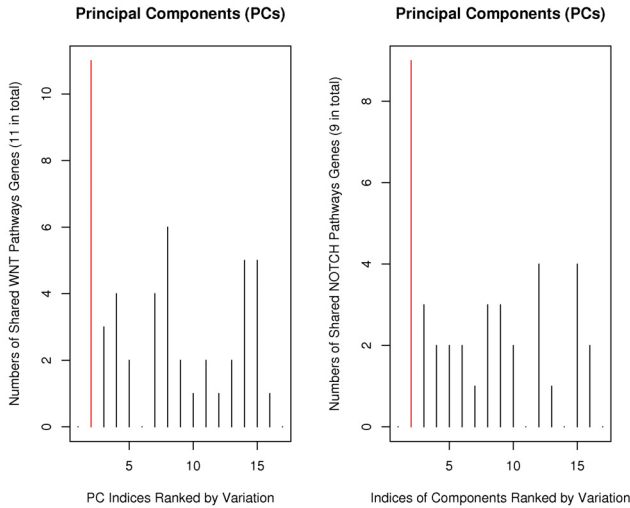
Figure 3
SVD analysis without splitting negative and positive PC's. WNT and NOTCH genes are maximally enriched (P-value: E-06) in the second PC (red lines), not the leading PC.

its pattern out of the noisy microarray data using appropriate analysis techniques, i.e., SVD.

The main contribution of this work is that we proposed an optimization algorithm combining the strengths of gene set based analysis and iterative gene selection. The iterative fashion inspired from the gene shaving algorithm allows distilling desired gene cluster using prior knowledge, while the latter enables us to discover gene clusters of modest and concerted expression change. The PC's that define gene clusters group a series of tightly regulated genes ranked by variance over samples. The orthogonality as specified in SVD analysis indicates those gene clusters of different variation were regulated by orthogonal defined or undefined genomic influences (Table 1 of [11]).

Our method is particularly suitable for identifying gene clusters with modest and concerted expression change, therefore it is not limited to identify periodically expressed gene clusters. When there is no prior knowledge available, the optimization process can be done through optimizing the enrichment of interesting Gene Ontology (GO) vocabulary, for example, somitogenesis [GO:0001756]. The technique for testing enrichment of GO term is very similar to that was used here, also see review in [30]. A recursive dendrogram can be constructed as a foundation to generate overlapping gene clusters, from which the optimal clusters can be identified and retrieved according to the enrichment of the interesting GO term(s) [3].

Conclusion
Our algorithm represents one of the first clustering algorithms that is particularly designed to identify signaling pathways of low and concordant gene expression variation. The discriminating power is achieved by manufacturing a principal component enriched by the prior knowledge.

Methods
Singular Value Decomposition
Assume the gene expression data is in the matrix format $X_{p \times n}$, where rows (p) correspond to genes and columns (n) correspond to conditions under which gene expression abundance were interrogated. Singular value decomposition (SVD) of the rectangular matrix X can be expressed as follows:

$$X_{p\times n} = U_{p\times n}S_{n\times n}V^T_{n\times n}, \qquad (1)$$

where $U_{p \times n}$ is the gene coefficient, and U_{ij} is the contribution of i_{th}, $i = 1, ..., p$, gene to the j_{th}, $j = 1, ..., n$, PC. If we correspond each U_j to a genomic influence j, then U_{ij} defines how much the gene i is regulated by the genomic influence j. $S_{n \times n}$ is the singular value matrix, where the diagonal contains list of singular values, and the magnitude of singular values corresponds to percentage of variation explained by each PC. $V^T_{n\times n}$ stores PC's [9,10]. We then separated positive PC's from negative PC's according to the signs of entries in $U_{p \times n}$, i.e.,

$$X_{p\times n} = X^+_{p\times n} + X^-_{p\times n}. \qquad (2)$$

Refer to supplemental figure 1 for a schematic illustration of the procedure. As shown in later data analysis examples, the separation operation is the key to enhance the prior knowledge enrichment level and to differentiate between antiphased WNT and NOTCH clusters.

Testing gene coefficients
Smaller fraction numbers of U_{ij} may indicate the contribution of i_{th} gene to j_{th} PC is negligible. We used a cut-off value that was originally used in [10] to test the vanishing of U_{ij} (similar to a 3σ statistical significance):

$$U_{ij} = \begin{cases} U_{ij} & \text{for } |U_{ij}| > \frac{p}{\sqrt{n}}, \\ 0 & \text{for } |U_{ij}| \leq \frac{p}{\sqrt{n}}. \end{cases} \qquad (3)$$

Each element in $X^+_{p\times n}$ and $X^-_{p\times n}$ is compared to the value $\frac{p}{\sqrt{n}}$, where n is the number of genes and p is a weight fac-
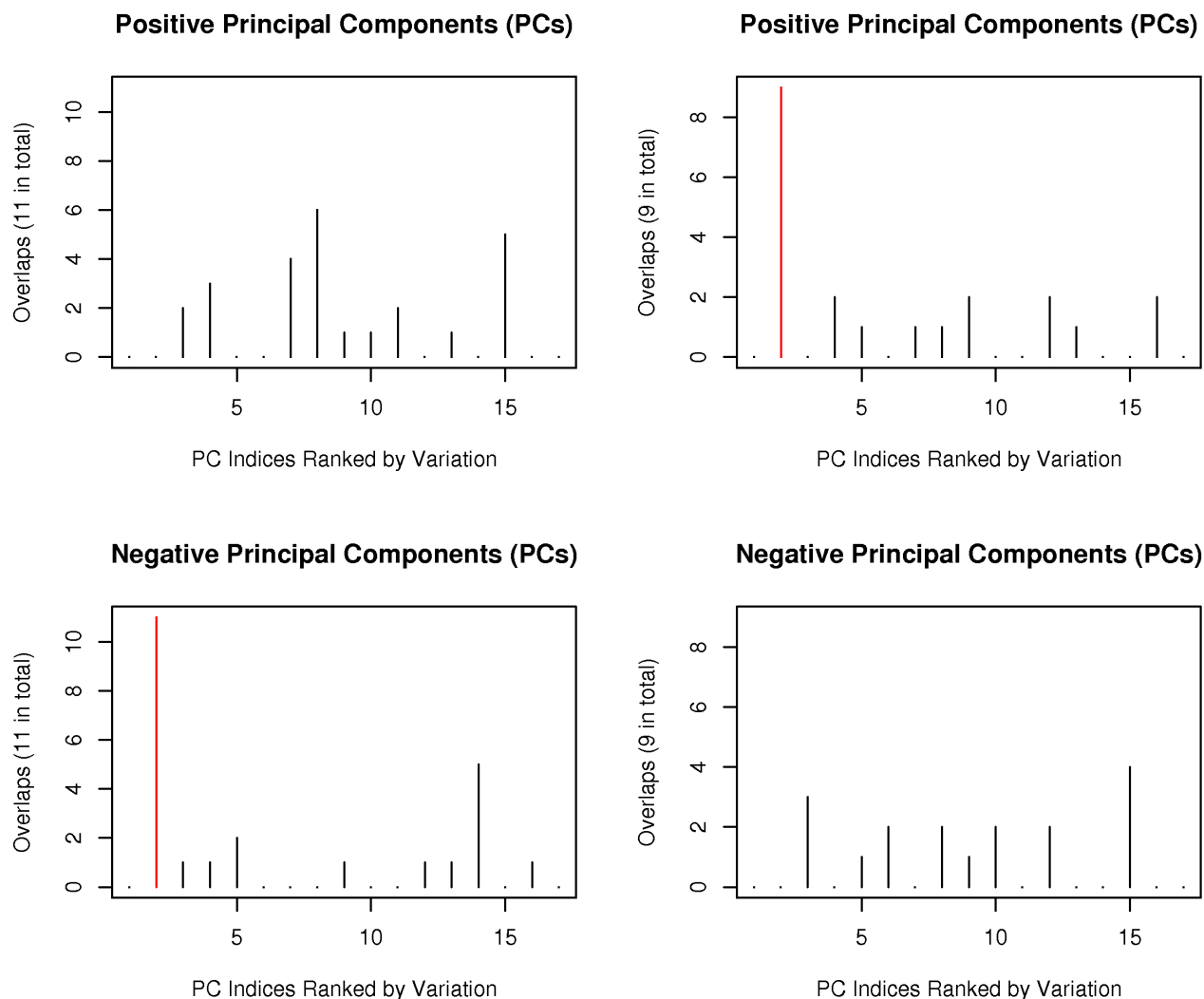
Figure 4
SVD analysis with splitting negative and positive PC's. Further, WNT and NOTCH genes are maximally enriched (P-value: E-10) in the second negative PC and the second positive PC (red lines), and the level of enrichment is dramatically increased because the sizes of negative and positive PC's decrease.

tor whose recommended value is 3. If the magnitude of the element in $X^{+}_{p \times n}$ and $X^{-}_{p \times n}$ is greater than $\frac{p}{\sqrt{n}}$, the corresponding gene is determined to contribute significantly to the PC's. Alternatively the list of genes that are significantly up-regulated or down-regulated by the underlying genomic influence corresponding to each PC.

Enrichment test
For each PC j, suppose there is a gene set K of k genes that U_{ij} is not 0, and for a biological pathway, suppose there is a prior knowledge gene set M of m genes in known in the

pathway. Also assume there are n genes NOT in the pathway, and x is the number of common genes shared by K and M. The probability of observing exactly x common genes is:

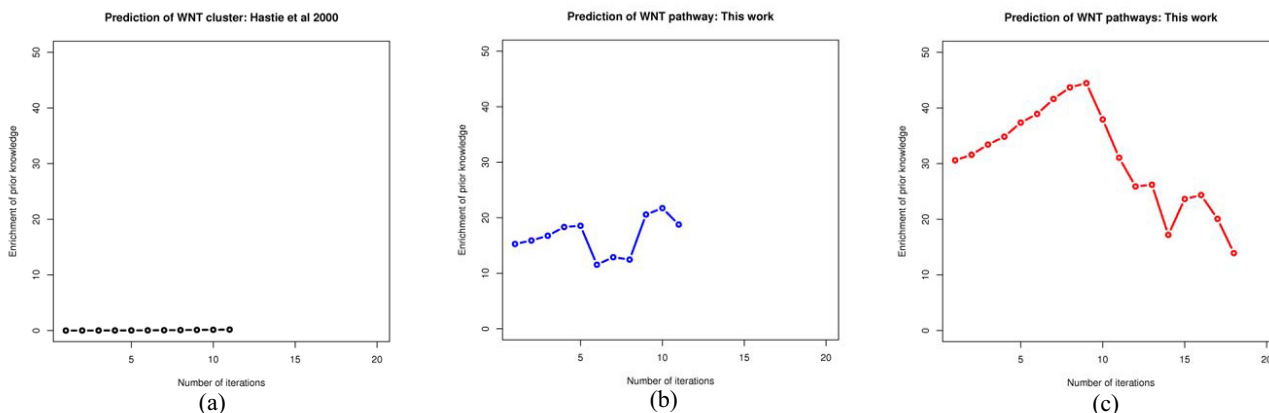$$P(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}. \quad (4)$$

Figure 5
Algorithm comparisons. Horizontal axis represents the number of iterations in both upper or lower panels. The vertical axis of the upper panel corresponds to the *-log2P*-value of the enrichment of prior knowledge. The vertical axis of the lower panel corresponds to the number of genes in the cluster (upper) and size of the cluster (lower). (a) The performance of the original gene shaving algorithm gauged by prior knowledge enrichment over iterations [4]. (b) The performance of our semi-supervised gene shaving algorithm without splitting positive and negative PC's. (c) The performance of our semi-supervised gene shaving algorithm with splitting positive and negative PC's.

In order to estimate the probability of observing x common genes or more is purely due to chance, we test the following one-sided hypothesis:

$$H_0 : O_1 = O_2 \text{ versus } O_1 \geq O_2, \qquad (5)$$

where O_1 is a parameter corresponding to the probability of genes in the prior knowledge belonging to the PC, and O_2 is a parameter corresponding to the probability of genes not in prior knowledge belonging to the PC. Under H_0, the test statistic x follows a hypergeometric distribution with known parameters m, n and k.

The P-value is then defined as the probability of observing x or more overlaps given H_0 is true. Therefore, it is calculated as follows:

$$
\begin{aligned}
P_V &= P(X \leq x) \\
&= 1 - P(X < x) \\
&= 1 - \sum_{o=1}^{x-1} P(X = o) \\
&= 1 - \sum_{o=1}^{x-1} \frac{\binom{m}{o}\binom{n}{k-o}}{\binom{m+n}{k}}
\end{aligned}
\qquad (6)
$$

Semi-supervised gene shaving algorithm
1: Start with the centered data matrix X that each row has zero mean

2: **while** TRUE **do**

3:    Singular   value   decomposition $X_{p \times n} = U_{p \times n}^{+} S_{n \times n} V_{n \times n}^{T} + U_{p \times n}^{-} S_{n \times n} V_{n \times n}^{T}$

4: **for all** column of $U_{p \times n}^{+}$ and $U_{p \times n}^{-}$ **do**

5: if column elements are greater than a cut-off **then**

6:   NO change

7: **else**

8:   Set to 0

9: **end if**

10: **end for**

11: **for all** Gene sets correspond to each columns **do**

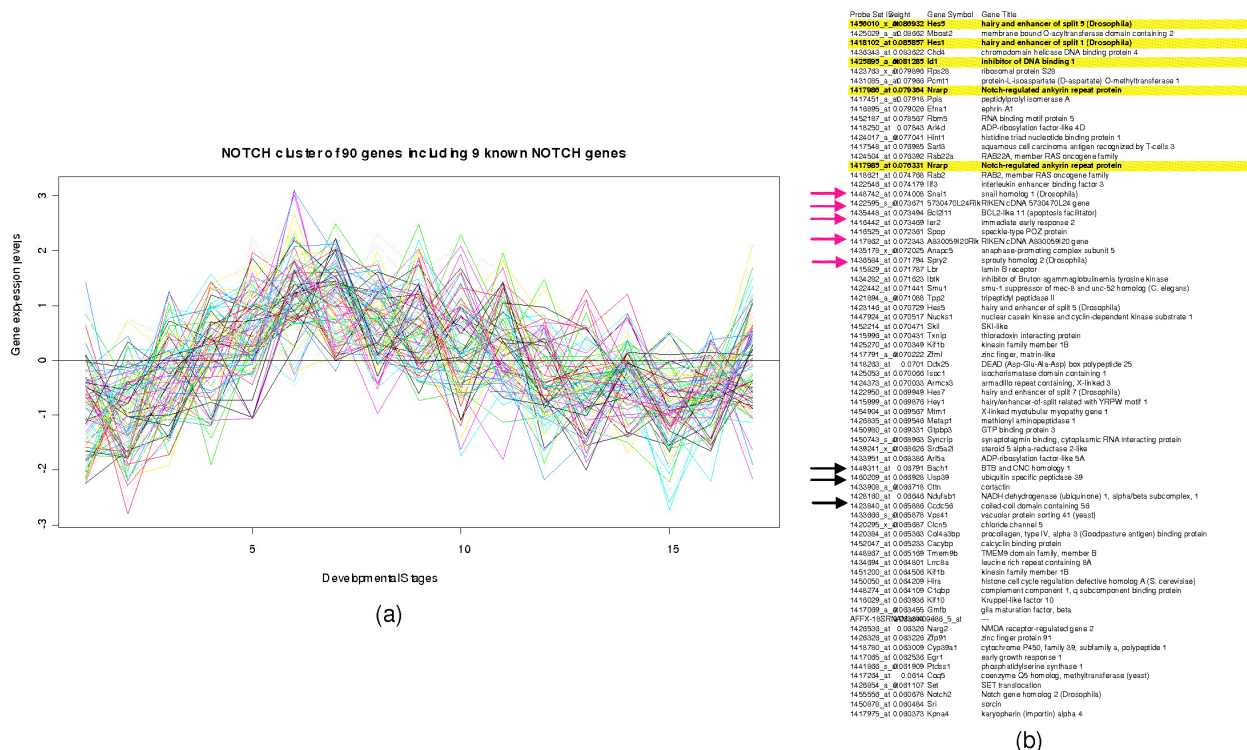12:   Test enrichment of prior knowledge in each gene set

13: **end for**

Figure 6
The predicted NOTCH cluster. Highlighted genes are prior knowledge. Genes that are pointed by red arrows correspond to experimentally validated NOTCH genes, and genes pointed by blue arrows correspond to potentially interesting genes by expert opinion and literature search. The whole list of prior knowledge and prediction are available in supplemental tables.

14:     **if** Two or more columns that are most enriched with prior knowledge exist **then**

15:       Break

16:     **else**

17:       Retrieve the best PC that are most enriched by prior knowledge

18:     **end if**

19:       Sort genes according to absolute correlation with the best PC

20:       Discard α% least correlated genes (α = 10% followed from [4])

21:       Assign the reduced data matrix to X

22: **end while**

23: Trace-back to retrieve the best gene cluster

As shown in the above Algorithm and Figure 1, the algorithm iterates until there are two or more most enriched PC's coexisting as defined by prior knowledge. The iterations stop here since we don't yet know a good way to further reduce the size of the cluster. Inconsiderate reduction might cause a loss of important genes. There are two ways of tracing back to retrieve the best gene cluster. One is to find the smallest cluster containing all prior knowledge, another is to find the cluster in which the enrichment of prior knowledge optimized. We chose the latter because it does not rely on the assumption that all prior knowledge need to be accurate. In fact, each gene coefficient can be used to measure the relative importance of genes in forming the cluster pattern. Genes in prior knowledge that help shaping out patterns receive higher weight, otherwise receive lower weight.

Stability analysis of gene clusters – a jackknife approach
Jackknife approach, e.g. "leave-one-out", is a resampling approach that is frequently used to access the stability of

Table 1: Biological function enrichment analysis and transcription factor association analysis.[]

Gene Set	Size	GO Annotation	Transcription Factors
WNT	45	embryonic development (1.13E-04) cytosol (9.15E-06) cytosolic part (4.48E-08) iron ion binding (3.92E-06) tube development (3.86E-04) branching morphogenesis of a tube (9.88E-06) tube morphogenesis (7.26E-05) patterning of blood vessels (3.57E-05) embryonic pattern specification (1.11E-04) oxygen binding (6.89E-14) gas transport (4.60E-14) hemoglobin complex (1.12E-14)	MyoG_Myotubes (9.47E-03) [24] MyoD_Growing cells (1.99E-05) [24]
NOTCH	36	developmental maturation (3.86E-04) negative regulation of cell differentiation (3.01E-04) ectoderm development (1.91E-05) cell maturation (1.94E-04) tissue morphogenesis (1.12E-05) epidermis morphogenesis (2.00E-06) hair cell differentiation (5.26E-06) mechanoreceptor differentiation (7.56E-06) negative regulation of neuron differentiation (3.49E-06) regulation of neuron differentiation (3.93E-05) cell fate determination (9.65E-06) auditory receptor cell fate commitment (3.78E-08)	MyoG_Myotubes (9.47E-03) [24] MyoD_Growing cells (1.99E-05) [24]

The third column contains biological functions significantly enriched in WNT and NOTCH pathways, and the fourth column contains transcription factors significantly associated with WNT and NOTCH pathways. The analysis was done through the web-server of the Segal lab: [23]

an estimator such as enrichment studied here. Suppose we wish to estimate enrichment parameter (η) as a complicated statistic (T) of n genes in prior knowledge as well as $\mathcal{D}$,

$$\hat{\eta} = T(g_1, g_2, ..., g_{i-1}, g_i, g_{i+1}, ..., g_n, \mathcal{D}). \quad (7)$$

Let jth partial estimate of η be given by the estimate computed with gene i removed,

$$\hat{\eta}_j = T(g_1, g_2, ..., g_{i-1}, g_{i+1}, ..., g_n, \mathcal{D}). \quad (8)$$

The jackknife estimate of η is given by the average of the pseudovalues [31],

$$\eta^* = \frac{1}{\eta} \sum_{i=1}^{n} (n\eta - (n-1)\eta_j). \quad (9)$$

An approximate sampling error for $\hat{\eta}^*$ can be obtained as the following [31]:

$$Var(\eta^*) = \frac{Var(\eta_j^*)}{n} = \frac{\sum_{j=1}^{n}(\eta_j^* - \eta^*)^2}{n(n-1)}. \quad (10)$$

Likewise, an approximate (1 - α)% confidence interval is given by [31],

$$\eta^* \pm t_{\alpha/2, n-1} \sqrt{\frac{\sum_{j=1}^{n}(\eta_j^* - \eta^*)^2}{n(n-1)}}, \quad (11)$$

where $t_{\alpha/2, n-1}$ satisfies $Pr(t_n \geq t_{\alpha/2, n-1}) = \alpha$, with t_n denoting a t-distributed random variable with n degree of freedom.

Competing interests
The author declares that they have no competing interests.

Author's contributions
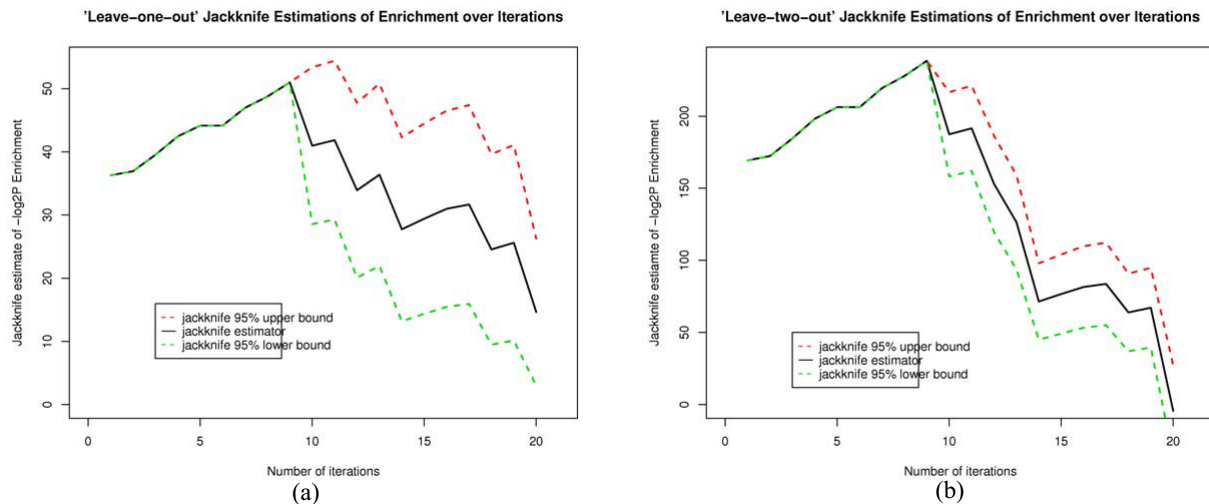DZ conceived and designed the method, analyzed data and drafted the manuscript.

Additional material

Additional file 1
Supplemental figures.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S1-S54-S1.doc]

Figure 7
Leave-one-out and Leave-two-out Jackknife estimations and confidence intervals of the enrichment. (a) Accessing cluster sensitivity to perturbation of prior knowledge using leave-one-out approach (b) Accessing cluster sensitivity to perturbation of prior knowledge using leave-two-out approach.

Acknowledgements

References

1. Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14587-15151.
2. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub T: **Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96(6):**2907-2912.
3. Zhu D, Hero A, Cheng H, Khanna R, Swaroop A: **Network constrained clustering for gene microarray data.** *Bioinformatics* 2005, **21(21):**4014-4020.
4. Hastie T, Tibshirani R, Eisen M, Alizadeh A, Levy R, Staudt L, Chan W, Botstein D, Brown P: **'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns.** *Genome Biology* 2000, **1(2):**.
5. Yeung K: **Model-based clustering and data transformations for gene expression data.** *Bioinformatics* 2006, **17(10):**977-087.
6. Gasch A, Eisen M: **Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.** *Genome Biology* 2002, **3(11):**.
7. Qin Z: **Clustering microarray gene expression data using weighted Chinese restaurant process.** *Bioinformatics* 2006, **22(16):**1988-1997.
8. Do K: **Applications of gene shaving and mixture models to cluster microarray gene expression data.** *Cancer Informatics* 2007, **2:**25-43.
9. Alter O: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97(18):**10101-10106.
10. Wall M, Dyck P, Brettin T: **SVDMAN – singular value decomposition analysis of microarray data.** *Bioinformatics* 2001, **17(6):**566-568.
11. Carter G, Rupp S, Fink G, Galitski T: **Disentangling information flow in the Ras-cAMP signaling network.** *Genome Research* 2006, **16:**520-526.
12. Lazzroni L: **Plaid models for gene expression data.** *Statistica Sinica* 2002, **12:**61-86.
13. Liang L, Mandal V, Lu Y, Kumar D: **MCM-test: a fuzzy-set-theory-based approach to differential analysis of gene pathways.** *BMC Bioinformatics* 2008, **9(Suppl 6):**S16.
14. Mootha V, Lindgren C, Eriksson K, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly M, Patterson N, Mesirov J, Golub T, Tamayo P, Spiegelman B, Lander E, Hirschhorn J, Altshuler D, Groop L: **PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nature Genetics* 2003, **34(3):**267-273.
15. Tian L, Greenberg S, Kong S, Altschuler J, Kohane I, Park P: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2005, **102(38):**13544-13549.
16. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub T, Lander E, Mesirov J: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102(43):**15545-15550.
17. Rustici G, Mata J, Kivinen K, Lio P, Penkett C, Burns G, Hayles J, Brazma A, Nurse P, Bahler J: **Periodic gene expression program of the fission yeast cell cycle.** *Nature Genetics* 2004, **36(8):**809-817.
18. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9(12):**3273-3297.
19. **PTC versus paired normal thyroid tissue** [http://www.sanger.ac.uk/PostGenomics/S_prombe/]
20. Dequeant M, Glynn E, Gaudenz K, Wahl M, Chen J, Mushegian A, Pourquie O: **A Complex Oscillating Network of Signaling Genes Underlies the Mouse Segmentation Clock.** *Science* 2006, **314(5805):**1595-1598.

21. Glynn E, Chen J, Mushegian A: **Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms.** *Bioinformatics* 2006, **22:**310-316.
22. **A Complex Oscillating Network of Signaling Genes Underlies the Mouse Segmentation Clock** [http://www.ebi.ac.uk/microarray-as/aer/#ae-browse/q=E-TABM-163[2]]
23. **Segal lab website** [http://genie.weizmann.ac.il/genomicaweb/enrichment/genesets.jsp]
24. Blais A, Tsikitis M, Acosta-Alvear D, Sharan R, Kluger Y, Dynlacht B: **An initial blueprint for myogenic differentiation.** *Gene & Development* 2005, **19(48):**553-569.
25. Cao Y, Kumar R, Bennett H, Charlotte A, Kooperberg C, Boyer L, Young R, Tapscott S: **Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters.** *The EMBO Journal* 2006, **25:**502-511.
26. Pittler S, Zhang Y, Chen S, Mears A, Zack D, Ren Z, Swain P, Yao S, Swaroop A, White J: **Incorporating gene functions as priors in model-based clustering of microarray gene expression data.** *Bioinformatics* 2006, **22(7):**795-801.
27. Larson P, Almasri E, Chen G, Dai Y: **A statistical method to incorporate biological knowledge for generating testable novel gene regulatory interactions from microarray experiments.** *BMC Bioinformatics* 2007, **8(317):**.
28. Tseng G: **Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data.** *Bioinformatics* 2007, **23(17):**2247-2255.
29. Tomphor J, Lu J, Kepler T: **Pathway level analysis of gene expression using singular value decomposition.** *BMC Bioinformatics* 2005, **6:**225.
30. Rivas E, Personnaz L: **Enrichment or depletion of a GO category within a class of genes: which test?** *Bioinformatics* 2006, **23(4):**401-407.
31. Manly B: *Randomization, Bootstrap and Monte Carlo Methods in Biology* Boca Raton: Chapman and Hall; 1997.