Time Course Analysis of Microarray Data for the Pathway of Reproductive Development in Female Rainbow Trout

Y. Liu[1], J. Verducci[1,2,*], I. Schultz[3], S. Hook[3], J. Nagler[4], G. Craciun[5], K. Sundling[6] and W. Hayton[7]

[1] *Department of Statistics, The Ohio State University, Columbus, OH, USA*

[2] *Mathematical Biosciences Institute, The Ohio State University, Columbus, OH, USA*

[3] *Battelle Pacific Northwest National Laboratory, Marine Sciences Laboratory, Sequim, WA, USA*

[4] *Department of Biological Sciences and Center for Reproductive Biology, University of Idaho, Moscow, ID, USA*

[5] *Departments of Mathematics and Biomolecular Chemistry, University of Wisconsin, Madison, WI, USA*

[6] *Programs in Biotechnology, Biophysics and Medical Sciences, University of Wisconsin, Madison, WI, USA*

[7] *Division of Pharmaceutics, The Ohio State University, Columbus, OH, USA*

Received 29 August 2008; revised 21 May 2009; accepted 28 May 2009
DOI:10.1002/sam.10047
Published online 13 August 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: New statistical procedures are introduced to investigate gene activity in support of the hypothalamus–pituitary–gonad–liver signaling network that provides the neuroendocrine regulation for reproduction in female, oviparous fishes. The methods include Shrunken Centroid Ordering by Orthogonal Projections (SCOOP) and a robust encoding of B-splines via Friedman's Generalized Elastic Net (GEN). SCOOP orders genes according to a novel criterion that balances discriminatory and correlative information. It is particularly useful in the present context where genes are only partially annotated, relevant networks are not known *a priori*, and the sample size is adequate for finding natural clusters. In this application, microarray measurements of gene expression were made in the pituitary, liver, and ovary of female rainbow trout *(Oncorhynchus mykiss)* over the course of their 1 year spawning cycle, and new methods were developed to detect systematic changes in potential gene networks. B-splines were fitted to gently smooth the estimates of expression versus time and provide a common framework for analysis. Unlike other methods, SCOOP selected not only the genes whose curves vary the most over time, but also genes closely correlated with these. This tended to recognize genes that may be part of an active network, but whose expressions undergo more modest fluctuations. To compensate for the high degree of uncertainty in fitting B-splines to individual genes, GEN methods were used to provide robust fits, and these were summarized through a novel GEN-transform as variable importance measures (VIMs). Clustering of genes via VIMs produced much more stable results than directly via B-spline coefficients, and the mean time course pattern of each cluster provided biologists with a reliable summary from which to interpret systematic patterns. Ultimately, the genes selected by SCOOP and clustered though the GEN-transform strongly suggested supportive pathways involving immunology, muscle contraction, reproduction, protein transport, metabolism, and reduction/oxidation. © 2009 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 2: 192–208, 2009

Keywords: B-splines; FSH; gonadotropins; generalized elastic net; generalized path seeking; LH; SCOOP

1. INTRODUCTION

The goal of this study was to identify networks of genes that support reproduction in female rainbow trout, based on microarray measurements of gene expression taken on different organs at strategically chosen time points and summarized by key time intervals (epochs) over the course of the annual cycle. No *a priori* knowledge of the networks was assumed. There are several novel aspects to the approach presented here: (1) use of within-epoch covariance to help elicit network structure; (2) balance of

Additional Supporting Information may be found in the online version of this article.

Correspondence to: J. Verducci (jsv@stat.ohio-state.edu) within-epoch and between-epoch covariance information in selecting genes that either vary substantially over the annual reproductive cycle or are strongly associated with ones that do; (3) use of transformations to compensate for the uncertainty in fitted time course patterns of genes when attempting to cluster these patterns.

The ultimate test of any new procedure is what it contributes to the scientific research to which it is applied. The networks identified through the use of our novel procedures, in particular the Shrunken Centroid Ordering by Orthogonal Projections (SCOOP) procedure for gene selection, make good biological sense and provide new insights, whereas the genes selected by the popular statistical procedure Extraction of Differential Gene Expression (EDGE) [1], though individually significant, failed to identify coherent networks.

1.1. Biological Background and Research Question

Although there is considerable information on the morphological and physiological changes that accompany a cycle of reproduction in numerous fishes, the application of molecular tools to study fish reproduction is just beginning [2–4]. The vast majority of fishes are oviparous (egg laying with external fertilization) and gonochoristic (sexually reproducing with distinct sexes), and members of the Family *Salmonidae* fall into this group. A common iteroparous (multiple spawning episodes) representative of this family, for which the basic reproductive biology is very well known, is the rainbow trout *(Oncorhynchus mykiss)*. Rainbow trout are medium-sized fish as adults (weight = 1–3 kg; ~45 cm total length) and females produce 2000–3000 eggs/kg body weight at the culmination of each annual breeding cycle [5]. All mature eggs are released for external fertilization during one or more closely spaced spawning events. The female reproductive cycle is coordinated by the hypothalamic region of the brain, which releases gonadotropin-releasing hormone (GnRH) via neurons that extend into the pituitary gland at the base of the brain [6]. The pituitary gland releases two gonadotropins, follicle-stimulating hormone (FSH) early in the cycle and luteinizing hormone (LH) at the end that serve to coordinate much of the development occurring in the gonads [7]. Besides providing the specialized environment in which the oocytes develop to form eggs, the ovaries of the female synthesize several hormones that have important reproductive roles both within and external to the ovaries. The ovaries produce estrogens, mainly 17β- estradiol in the rainbow trout, during the phase of active ovarian growth, and progestins, principally 17α, 20β-dihydroxy- 4-pregene-3-one, at the end of the cycle that are necessary for final oocyte maturation [8].

In oviparous female fishes, an additional organ of critical importance is the liver. The liver is the site of vitellogenin synthesis, which is the egg yolk precursor protein responsible for the bulk of the growth observed in the ovary [9]. The liver is stimulated to produce vitellogenin by estrogen released from the ovary and is therefore important to be considered in the coordination of oogenesis in oviparous female fishes. Collectively, these organs make up the hypothalamus–pituitary–gonad–liver (HPGL) signaling axis or network that provides the neuroendocrine regulation for reproduction in female fishes (Fig. 1).

The purpose of this study was to use microarray data from a time course study of the female rainbow trout (*Oncorhynchus mykiss*) to enhance our understanding of the gene pathways associated with the complete (annual) reproductive cycle. Although several facets of the HPGL signaling network are understood, it is naïve to expect that other important factors are not directly or indirectly involved. To address this objective, the change in gene expression level in three key organs (pituitary, ovary, and liver) that make up the HPGL signaling network were analyzed.

1.2. Previous Approaches to Analysis of Time Course Microarray Data

Bar-Joseph *et al.* [10] introduced spline smoothing of time course microarray data using a random effects model

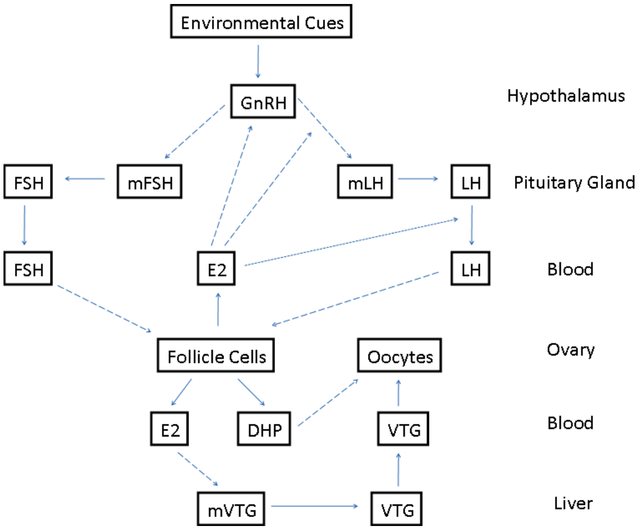

Fig. 1 The pathway of environmental control over the hypothalamus–pituitary–gonad–liver signaling network in female oviparous fishes. The different tissue compartments are indicated on the right hand side. Solid lines represent synthesis/ release/uptake, dashed lines represent positive endocrine feedback, and dotted lines represent negative endocrine feedback. (GnRH: gonadotropin-releasing hormone, FSH: follicle-stimulating hormone, mFSH: follicle-stimulating hormone mRNA, LH: luteinizing hormone, mLH: luteinizing hormone mRNA, E2: 17β-estradiol, DHP: 17,20β- dihydroxy- 4-pregene-3-one, VTG: vitellogenin, and mVTG: vitellogenin mRNA).

(assuming all genes have the same covariance matrix over time) and E−M (expectation−maximization) method to get stable profiles for mean expression of (known or unknown) sets of genes. The inclusion of a fixed number of classes offsets the difficulties associated with clustering many individual, highly variable, curves. Shortcomings with use of this approach here are that it requires a high rate of sampling over the time course, and the specification of a fixed number of classes. It is also computationally intensive in dealing with the uncertainty in curve fitting. The examples were limited to 800 genes; here we have 16 000 genes and need faster methods to screen all genes. Nevertheless, spline smoothing of time course microarray data is a fundamentally sound and important step in analyzing any time course data sampled strategically enough to capture periods of change.

Storey *et al.* [1] developed univariate F-statistics for coefficients of spline basis functions and used false discovery rate (FDR) methods to adjust for multiple comparisons, which is the basis for the popular EDGE software. Difficulties here include the fact that univariate methods generally focus on finding just a few key, isolated genes. In our application, the changes in gene expression over the spawning cycle are extensive and systemic. Nevertheless, EDGE may be useful for identifying particularly significant genes.

Tai and Speed [11,12] generalized the Hotelling T^2 test statistic to the time dependent, small number of replicates case, and developed a multivariate (over time points) empirical Bayes (MB) statistic to rank genes for differential expression between conditions on the basis of time course data. This work is incorporated in the bioconductor "timecourse" package available for R software (http://www.bioconductor.org/packages/2.3/bioc/vignettes/timecourse/inst/doc/timecourse.pdf). This method makes no use of the actual time scale, and each gene is modeled individually. Because the time scale is extremely important here, "time course" was not used. Importantly, though, the method demonstrates that gene ranking can be very sensitive to estimation, an issue discussed in the last section of this paper.

Nueda *et al.* [13] recommended analysis of variance-simultaneous component analysis (ANOVA-SCA) to take advantage of the structured experimental design. This entails a basic ANOVA approach, but uses principal component analysis (PCA) simultaneously on various covariance matrices to reduce noise due to high dimensionality. Difficulties here include irregular time sampling which is different for pituitary, liver, and ovaries; also no use is made of the actual times and so the method lacks direct time smoothing. Nevertheless, the idea of component specific dimension reduction via PCA is inspiring. In particular, we also use ideas from PCA both to reduce noise and to establish a suitable stopping criterion for our suggested selection of relevant genes. However, we use a novel idea of shrinking group means toward an *enriched discriminate space*, which is the direct sum of the linear spaces encoding gene covariation, both within and between epochs.

Telesca *et al.* [14] constructed a three-stage hierarchical model. At the first stage is a common shape function formed from B-splines; at the second stage, individual gene variations are modeled by individual levels, amplitudes, and time transformations; at the third stage, amplitudes are restricted to fall into one of three classes (+, 0, −), depending on the sign of their weighting on the common shape function. A fully Bayesian analysis is used and additional information on known (regulator, response) gene pairs, especially in conjunction with time transformations, leads to identification of likely transcription factor/response mechanisms. Difficulties here include the requirement of a high sampling rate and the assumption of only one cluster amid noisy genes. We have not attempted to generalize this type of analysis to our data.

1.3. Outline of the Paper

The next section describes the overall method of analysis based on new and adapted informatics tools. Section 3 provides details of the two new methods SCOOP and GEN-transform. Sections 4 and 5 contain the results and discussion.

2. OVERALL METHOD OF ANALYSIS

2.1. Fish and Microarray Analysis

A group of 3-year-old female rainbow trout were selected from a population maintained at a commercial rainbow trout supplier (Troutlodge Inc., Sumner, WA, USA). These fish were transported to the Battelle Marine Sciences Facility (Sequim, WA, USA) immediately after they had spawned for the second time and were placed in a tank with flowing freshwater (12°C) and natural photoperiod. Three fish were

Table 1. Number of fish sampled during the spawning cycle.

Cycle day	8	57	81	92	100	107	130	161	190	220	249	260	282	301	316	317	330
Pituitary	0	3	3	2	1	0	0	0	3	0	3	3	2	3	3	0	3
Ovary	3	3	3	0	0	3	0	3	1	2	0	3	3	3	2	1	0
Liver	0	3	1	2	0	0	1	0	3	3	0	0	3	3	3	0	3

harvested at various times over the course of 1 year to cover a complete reproductive cycle, as depicted in Table 1. Tissue samples from the pituitary, ovary, and liver were immediately placed in RNA*later*™ solution (Ambion, Austin, TX, USA), gradually chilled to 4°C and eventually frozen at −80°C for storage.

A microarray designed (Batch number: HL002 - IB009) for salmonid fishes was purchased from GRASP (http://web.uvic.ca/grasp/microarray/array.html) and used as the platform for this study. This microarray has 16 000 cDNA or ESTs from several salmonid fishes, including Atlantic salmon (*Salmo salar*), rainbow trout (*Oncorhynchus mykiss*), and Chinook salmon (*Oncorhynchus tshawytscha*). Pooled cycle day 1 (defined as the day after spawning) samples from the corresponding organ were referenced as controls in a two-channel format on the microarrays. RNA was transcribed into cDNA and labeled via aminoallyl technique using Invitrogen's Superscript™ Indirect cDNA labeling kit. Reference and experimental cDNA were labeled with Alexa fluor 555 or Alexa fluor 647 fluorescent dyes, respectively. A split control experiment (where reference was labeled with Cy3 and Cy5 and then combined for hybridization) was performed to assess the dye bias [15]. Both control and treatment samples were adjusted according to the concentrations of cDNA. The volume was then further reduced. The addition of 20 g herring sperm DNA prevented nonspecific hybridization in a modified hybridization buffer (50% formamide, 40% 20 × saline-sodium citrate (SSC), 9% Denhardt's solution, 1% sodium dodecyl sulfate (SDS)). The hybridization was carried out at 45°C for 16 h. After hybridization, microarrays were rinsed in SSC/SDS buffer to remove unbound or nonspecifically bound cDNAs. Subsequently, the microarray was quantified with a Perkin Elmer ScanArray Express imaging system with varied laser power and photo-multiplier tube (PMT) gain to equalize the fluorescence between the channels thus minimizing the possibility of oversaturation of the signal. The expression data were acquired with the ScanArray Express Software with the gene annotation list (GAL) file defined in http://web.uvic.ca/grasp/microarray/hl-ib_data.html. The median of the fluorescence intensity with background subtracted was further adjusted for bias using the LOWESS method [16], which is available through the website (http://stat-www.berkeley.edu/users/terry/zarray/Html/normspie.html). Additional calculations were carried out using the R software package (http://www.r-project.org/). The intensity was converted to a base 2 log scale with zero substituting log negative values. The control (Day 1) log intensity was subtracted from the experimental log intensity to provide normalized expression values for all microarrays. Average expression was calculated for each gene with multiple probes.

2.2. Analysis of Microarray Time Course Data

The complexity of signaling pathways suggested a global approach to identifying genes involved in the reproductive cycle of female rainbow trout. We adopted a five step approach: (1) *Smooth* data using B-splines; (2) *Screen* genes differentially expressed over time in all three organs using EDGE and SCOOP procedures, resulting in a manageable number of genes; (3) *Transform* information about curve fitting to *variable importance measures* using a Generalized Elastic Net; (4) *Cluster* genes using agglomerative and *k*-means methods on variable importance measures; (5) *Identify* the potential pathways of the differentiated sets of genes.

Natural B-splines are piecewise cubic polynomials, smoothly connected at knot points, which can describe the expression patterns as smooth curves over time. Each fitted curve is a weighted sum of basis functions, each of which roughly represents a short time period, referred to here as an epoch. B-splines afford a number of advantages. They make use of actual times at which samples are taken, avoiding the need to group or categorize these; easily accommodate different sampling schedules for different organs; reduce noise, assuming only smooth changes in expression over time; serve as local moving averages; and closely approximate all mean curves, given enough knots.

As a first step, we adapted the B-spline approach of Bar-Joseph *et al.* [10], using slightly different sets of basis functions in each organ because there were organ-specific strategic sampling schedules and we did not wish to overly bias estimates of the gene expression curves. For each organ (pituitary, liver, or ovary), we took second differences in the sequence of expressions for each gene, and estimated zero second derivatives as occurring midway between two time points where there was a change in the sign of the second differences. From a frequency table of all such estimated points of inflection, we chose the seven most frequent time points to be interior knots for all the splines used to fit gene curves for that organ. This produced a total of nine basis functions to summarize the time curves of gene expression measured at varying numbers of time points in each organ: pituitary (11 time points); ovary (12 time points), and liver (10 time points). Figure 2(a) depicts the basis functions used for pituitary expressions; Fig. 2(b) illustrates the individual measurements, piecewise linear pattern for mean expression, and B-spline smoothed expression curve of a particular pituitary gene.

Because we began with the full set of 16 000 genes, instead of modeling common class curves as in Bar-Joseph *et al.* [10] at this stage, we simply fitted individual curves, and used a combination of observed expression and fitted curves to make gene selections. This was accomplished by augmenting each individual fish measurement at a particular time with the observed mean values at other time points.

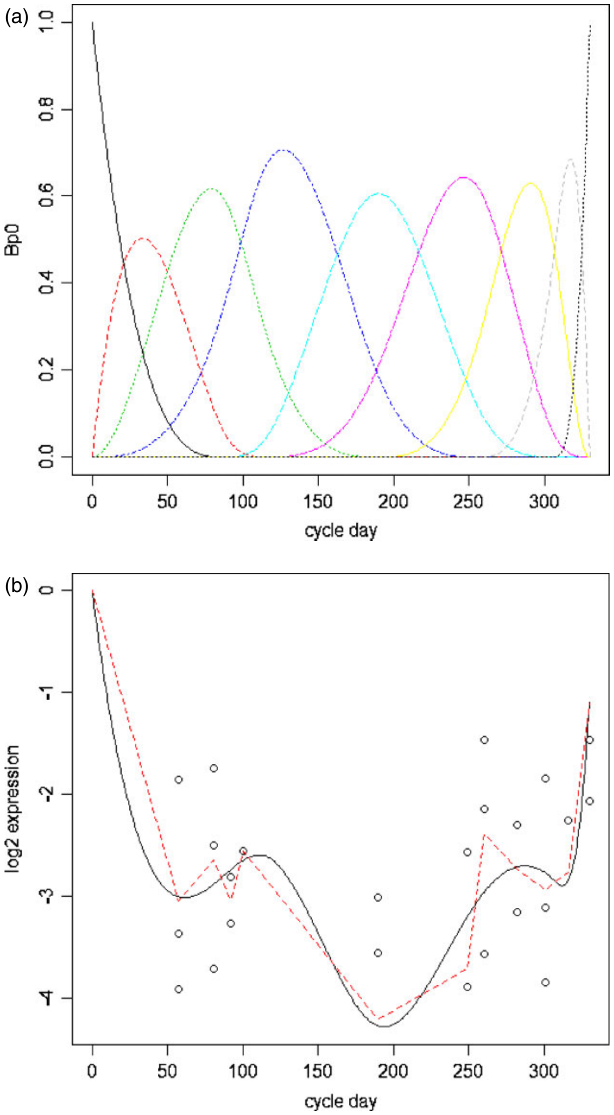

Fig. 2 (a) The B-spline basis functions for the pituitary of female rainbow trout over the course of a complete reproductive cycle. (b) Mean and B-spline smoothed time curves for gene CB49152 expression in the female rainbow trout pituitary over the reproductive cycle.

The resulting data set, with one "time series" per fish per gene (fish/gene) was used as input to the gene selecting procedures EDGE and SCOOP.

This preparation for the second step afforded two different ways of selecting genes whose expression changes systematically over the spawning cycle. The EDGE method [1] estimated the coefficients of a B-spline function to each individual fish/gene, and computed an F-statistic to calculate the p-value for each different gene. Alternatively, the estimated coefficients of a B-spline function to each individual fish/gene were input into the SCOOP procedure, discussed fully in the next section. SCOOP calculated the eigenvectors of the between- and within-epoch covariance matrices, and projected the data in a direction orthogonal to these. For each gene g, its importance was calculated as the maximum (over epochs) distance λ_g needed for an epoch mean to equal the overall mean. Genes were ranked separately according to the EDGE p-value and the SCOOP importance measure λ_g. For SCOOP, the cutoff was chosen by repeating the procedure, dropping the eigenvectors with smallest corresponding eigenvalues (explaining a total of less than 20% of the variation in each covariance matrix), and keeping all the top ranking genes up to the point where the two SCOOP rankings diverged. For EDGE, we first used the recommended rule of keeping the estimated FDR less than 10%. Under this criterion, EDGE selected 55 ovarian genes, 67 hepatic genes but no pituitary genes. We also extended the range to match the number (193) of genes selected by SCOOP.

Before attempting a clustering of curves for the selected genes, it is important to account for the variability with which they are estimated. For some genes, the observed expressions at collection times may be represented by many different splines almost equally well, leading to large standard errors and/or confounding for the coefficients of the B-splines. Thus, in the third step, instead of using fixed estimates for the coefficients, a sequence of estimates obtained under a penalized fitting procedure was transformed into a measure of variable (here a B-spline basis function) importance for the selected genes. The method, GEN-transform, uses Friedman's Generalized Elastic Net (GEN) to shrink coefficients toward 0, as described in the 2008 paper "Fast Sparse Regression and Classification." (http://www-stat.stanford.edu/~jhf/ftp/GPSpaper.pdf/). A particular, nonconvex, penalty for GEN separated variable effects more quickly than the traditional least absolute shrinkage and selection operator (LASSO) [6]. The amount of shrinkage needed to pin a coefficient at 0 is a measure of variable importance. The GEN-transform value we used in the clustering was a monotone function of this shrinkage. Thus, for each gene, the importance of each B-spline basis function, which represented a weighted time interval, was measured by its GEN-transform value. Each GEN-transform value had a direct interpretation in terms of the certainty with which there was a change in gene expression during the epoch associated with its B-spline basis function.

In step 4, agglomerative and k-means clustering were applied to the GEN-transform values to produce more stable clusters than if they had been applied to the original B-spline coefficients. We used the "agnes" procedure from the R package "cluster" to decide on four clusters, and then used k-means from the same package to refine the sorting into the four clusters. In step 5, a partial annotation table from Gene Ontology enabled interpretation of the functional roles suggested by the clustering. All together the analyses led to interpretable sets of genes that were differentially expressed in the female rainbow trout ovary, pituitary, and liver over the course of a reproductive cycle. The novel, bioinformatic approaches, SCOOP and GEN-transform described in the next two sections, are key tools enabling scientists to infer pathways that could be used to inform the reproductive biology of this and other oviparous fishes.

3. DETAILS OF METHODS

3.1. EDGE

EDGE uses B-spline smoothing and the time course for the expression of each gene i is assumed to follow the linear model:

$$\mu_{i(t)} = b_{i0} + b_i^T s(t) \tag{1}$$

where $\mu_{i(t)}$ is the mean expression of gene i at time t; b_{i0} is the intercept at time 0; $\boldsymbol{s(t)} = [s_1(t), \ldots, s_k(t)]^T$ is a vector containing the k basis functions; $\boldsymbol{b}_i^T = [b_1, \ldots, b_k]$ are the coefficients of the basis functions.

The null hypothesis to be tested is:

$$H_0 : b_i = 0 \text{ for all } i = 1, \ldots, k, \text{ versus} \tag{2}$$

$$H_A : b_i \neq 0 \qquad \text{for some } i \tag{3}$$

The p-value is based on an F-test. Genes were selected based on a Q-value limiting the FDR and the suggested cutoff Q-value is 0.10. Although some of the genes selected by EDGE are quite informative (see Section 4), the EDGE procedure suffers from several drawbacks. The basic deficiency is that EDGE evaluates each gene separately, so no connection is made among the different genes that might be operating in a common pathway. This deficiency may eventually disappear when genes become fully annotated, but currently fewer than half of the gene targets on the GRASP microarray are annotated, and many of the annotations are incomplete. Another deficiency is that EDGE relies on the assumption that log expression is normally distributed; a cursory inspection of the marginal distributions suggests there are many genes for which this is not true. A third deficiency is that variances might be better estimated using an empirical Bayes procedure, such as developed by Tai and Speed [11,12]. For these reasons, we supplemented the EDGE results with those from the novel procedure SCOOP.

3.2. SCOOP

First, we introduce some definitions and notations:

(1) Between-epoch covariance matrix:
Let X_{gtf} denote the value of gene g at time t for fish f, with $g = 1, \ldots, p$; $t = 1, \ldots, n_T$; and $f = 1, \ldots, n_F(t)$. Let $\vec{\overline{X}}_t$ be the p dimensional vector of average gene value at time t; And define the average value vector over time as:

$$\vec{\overline{X}} = \frac{1}{n_T} \sum_t \vec{\overline{X}}_t$$

And define the average value (over fish) for gene g at time t as:

$$\overline{X}_{gt} = \frac{1}{n_F(t)} \sum_f X_{gtf}$$

The *between-epoch covariance matrix* is defined as the $p \times p$ matrix

$$\begin{aligned} S_b &= \frac{1}{p(n_T - 1)} \sum_t (\vec{\overline{X}}_t - \vec{\overline{X}})(\vec{\overline{X}}_t - \vec{\overline{X}})' \\ &= \frac{1}{p(n_T - 1)} [\overline{X} - (1/\sqrt{n_T}) \vec{\overline{X}}\, \vec{1}'] \\ &\quad \times [\overline{X} - (1/\sqrt{n_T}) \vec{\overline{X}}\, \vec{1}']' \end{aligned}$$

where $\vec{1}$ is an $n_T \times 1$ vector of ones and $\overline{X} = [\overline{X}_{gt}]$ be the $p \times n_T$ matrix of the average values of genes for time. In our example, this matrix describes the variation of gene expression over different time intervals (epochs, represented by B-spline basis functions) and summarizes covariation over time of the fitted average gene value in each epoch.

(2) Within-epoch covariance matrix:
Let $\vec{D}_{tf}$ be the deviation of the value of gene g from its mean value at time t; let $D_{gtf} = X_{gtf} - \overline{X}_{gt}$ be the p dimensional vector of these deviations; Define the $p \times p$ covariance matrices of gene values at time t as:

$$S_t = \frac{1}{n_F(t) - 1} \sum_f (\vec{D}_{tf})(\vec{D}_{tf})'$$

Define the $p \times p$ (pooled) within-epoch covariance matrix by

$$\begin{aligned} S_w &= \frac{1}{\sum [n_F(t) - 1]} \sum_t [n_F(t) - 1] S_t \\ &= (X - \overline{X}_*)(X - \overline{X}_*)' / \sum [n_F(t) - 1] \end{aligned}$$

where X is the $p \times n$ data matrix, with $n = \Sigma_t n_F(t)$ being the total number of observations (fish); and $\overline{X}_*$ is a $p \times n$ matrix whose g-th row contains the time means $\overline{X}_{gt}$ each repeated $n_F(t)$ times. This matrix averages the covariation of fitted gene expressions within each epoch by pooling the variation in each epoch interval over all such intervals.

(3) Singular value decomposition (SVD) for S_w and S_b: Note that S_b and S_w are both very large square matrices, with dimension p typically in the tens of thousands. A direct eigen decomposition can consume inordinate amounts of time and space, making the calculation impossible on most laptop computers. However, because $p \gg n_T$ and $p \gg n_F(t)$, both S_b and S_w are singular, with respective ranks r_b and r_w each much less than p. This enables a quick calculation of the eigenvectors (with positive eigenvalues) of S_b and S_w through the singular value decomposition (SVD):

$$\overline{X} - (1/\sqrt{n_T})\vec{\overline{X}}\ \vec{1}' = \Gamma_b \Lambda_b \Psi_b$$

where Γ_b has dimension p x r_b, Λ_b is diagonal with positive entries and Ψ_b has dimension r_b x n_T. The columns of Γ_b are orthonormal vectors, the eigenvectors of

$$S_b = \frac{\Gamma_b \Lambda_b^2 \Gamma_b'}{p(n_T - 1)}.$$

The eigenvectors Γ_w of S_w may be computed using the SVD in a similar fashion:

$$X - \overline{X}_* = \Gamma_w \Lambda_w \Psi_w$$

so that

$$S_w = \frac{\Gamma_w \Lambda_w^2 \Gamma_w'}{n - n_T}.$$

is the spectral decomposition of S_w.
In the setting where $p \gg n$, the singular value decomposition operates on a much smaller $p \times n$ matrix rather than a $p \times p$ matrix, thus boosting the calculation rate of SCOOP.

(4) Enriched discriminant space (EDS):
The EDS is defined as the span of the union of eigenvectors of S_b and S_w, and a projection operator P_{bw} onto the EDS may be constructed by a Gram-Schmidt orthogonalization and renormalization to get Γ_{bw} with orthonormal column vectors that span the EDS. Since the eigenvectors can be viewed as the basis functions that span the space preserving the variation of the original matrix, the union of the eigenvectors of S_b and S_w then preserves the variation of the original data matrix. The basic idea of the SCOOP method is to shrink data toward the EDS, eliminating genes while preserving information. The shrinking is done along directions that are orthogonal to the EDS, as follows:
The matrix $P_{bw} = (\Gamma_{bw})(\Gamma_{bw})'$ serves as a projection operator onto the EDS; and $I_p - P_{bw}$, with I_p being the $p \times p$ identity matrix, projects data vectors onto the orthogonal complement of the EDS. Because the EDS contains all relevant linear discriminant information, projections of data vectors along the hyperplane $I_p - P_{bw}$ tend to preserve this information. That is the key idea behind SCOOP.

(5) Geometric interpretation of SCOOP:
For each gene, SCOOP moves the mean at time t along a direction orthogonal to the EDS until the means at all times are the same. See Fig. 3 for a schematic representation. The figure depicts observations at two time points, labeled with a green "x" or a red "o"; the means are labeled with capitol letters, with the overall mean shaded in blue. All points lie in the EDS, which here is depicted as a line. SCOOP moves each of the means in (opposite) directions that are orthogonal to the EDS until all means have equal values on one of the (gene) coordinate axes. Let δ_{gt} be the minimum distance that the mean for gene g at time t must move to make all the means for g equal, which is denoted as discriminant information. Then $\delta_g = \max_t\{\delta_{gt}\}$ represents a distortion incurred by dropping variable g. SCOOP ranks genes according to their δ_g values, dropping those with the smallest values.

(6) Stopping criterion for SCOOP:
A natural question is when to stop dropping genes and accept the top group as likely to be involved in systematic time course patterns. This can be done graphically by plotting log $\log(\delta_g)$ values against their rank and looking for a change in slope. The

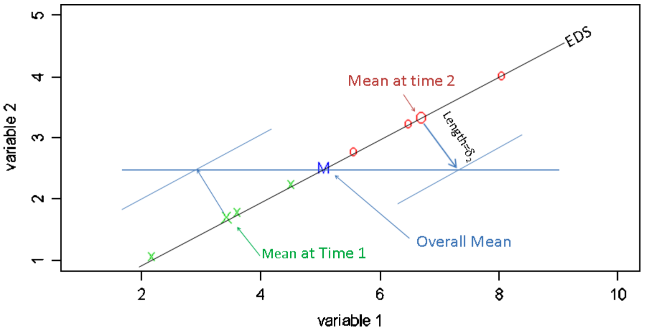

Fig. 3 Geometric illustration of the Shrunken Centroid Ordering by Orthogonal Projections method. The figure depicts observations at two time points, labeled with a green "x" or a red "o"; the means are labeled with capitol letters, with the overall mean shaded in blue. All points lie in the EDS (enriched discriminant space), which is depicted as a line.

asymptotic justification for this is that the extreme order statistics from a normal distribution tend to have Gumbel marginals. Alternatively, in the spirit of the SCA used by Nueda *et al.* [13], the following stopping rule seems quite reasonable: Rerun SCOOP using only the first K_{80} of the eigenvectors (number needed to explain 80% of variance) in S_b and S_w; stop when the rank order of the top genes changes. For the spawning cycle genes, both cutoff criteria serendipitously chose 193 genes.

(7) Simulations results of SCOOP:
To evaluate the performance of SCOOP in a simple context, a simulation test was performed. The simulated data were divided into two groups with 50 samples in each group. Data were generated from a normal distribution with 4000 variables, variances of which were all 1. The first 10 variables were simulated with correlation 0.75 between all pairs and a difference of 2 between group means; the second 10 variables were simulated with correlation 0.75 between all pairs and a difference of 1 between group means; the rest were simulated independently with the same mean of both groups. We called these 20 variables "informative." For the 100 simulations, the ranking of individual t-tests according to p-values was assessed. Seventy-three percent of the top 20 variables selected correctly corresponded to the informative variables; on average, the top 400 variables must be selected to include all 20 informative variables. In contrast, SCOOP was more efficient. Ninety-one percent of the selected top 20 variables were correct; and on average, only 200 variables were needed to ensure the inclusion of the 20 informative variables. The t-test was used for comparison because it could be viewed as the basis of many univariate variable selection methods. For example, EDGE may be viewed as an adapted time course version of the t-test (or F-test). This was one of the reasons why we compared our method with EDGE. Other simulation work covering the increased power of SCOOP over univariate approaches has been described in one author's (JV) presentation at the 2006 AMS-IMS-SIAM summer research conferences (http://www.cs.cmu.edu/∼lafferty/ml-stat2/).
In contrast to the popular EDGE software, which tests genes one at a time, SCOOP attempts to discover *sets* of correlated genes that change over time. EDGE and SCOOP methods provide distinct types of information about relevant genes. In particular the time course patterns of SCOOP-selected genes tend to be much more similar than those of the EDGE-selected genes, suggesting more natural clusters.

3.3. GEN-transform

We used a large set of B-spline basis functions to minimize bias when approximating the mean expression curves. For some curves, this representation is nearly redundant, and different linear combinations of basis functions can give rise to very similar B-spline curves. This leads to large standard errors and/or confounding in the least squares coefficients to describe the time course curve. A robust regularization method for estimating coefficient vector $\boldsymbol{b}$ proceeds as follows: Choose $\boldsymbol{b}$ to maximize:

$$\text{LogLikelihood } (\boldsymbol{b}) - \lambda^* \ \text{ Penalty } (\boldsymbol{b}),$$

where the LogLikelihood is constructed under the usual normal error linear model with mean expression modeled by (1). For $\lambda = 0$, this produces the least squares estimates. Under certain concave penalty functions, the sizes of the individual coefficients will decrease continuously (at different rates) to 0, as the regularization parameter λ increases. As $\lambda \to \infty$, all components of the maximizing coefficient vector go to 0. *If* λ_j denotes the largest value of λ at which the j-th coefficient differs from 0 ($j = 1, \ldots, 9$), then the importance of the j-th variable is robustly measured by λ_j. This transformation of the actual coefficients into variable importance measures (VIMs) $\{\lambda_j\}$ makes the representation more robust and the resulting clustering more stable.

A generalized path search (GPS), operates on a shrinkage scale τ inversely related to the penalty λ. GPS can be used to obtain shrinkage paths for all coefficients of B-splines, under squared error loss (Gaussian log likelihood) and the choice of GEN Penalty defined by

$$\text{Penalty } (\boldsymbol{b}) = \sum_{j=1}^{9} |\boldsymbol{b}_j|^{1/2}$$

These choices lead to continuous shrinkage paths for each coefficient, with relatively fast computation. For each gene, we record how many steps τ must be taken until each B-spline coefficient lifts off 0. The gene paths are summarized by these nine"rising times" $\tau = [\tau_1, \ldots, \tau_9]$, which are inversely related to the VIMs $[\lambda_1, \ldots, \lambda_9]$. The vector τ of rising times provides a stable summary of the form of gene expression curve. Figure 4 illustrates the shrinkage paths for the B-spline coefficients that characterize the gene expression curve for gene CB509494. Note that B-splines BS4 and BS5, although they have the largest coefficients when no shrinkage is used (right side limit not reached in this figure), are somewhat compensatory (coefficient estimates are negatively correlated), and actually are the least informative in describing the curve.

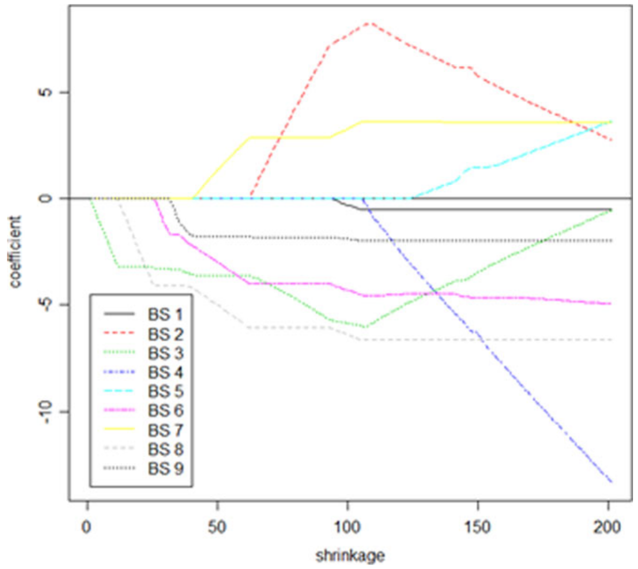

Fig. 4 The shrinkage paths of the B-spline coefficients for gene CB509494 in the pituitary of female rainbow trout over the reproductive cycle using the Generalized Elastic Net penalty. Coefficients are the B-spline coefficients. In total, there are nine basis functions for each gene labeled BS 1-9.

4. RESULTS

4.1. Eigenvalue Plots

Between- and within-epoch covariance matrices play an important role in the SCOOP procedure. Because these large matrices are constructed from relatively few observations, it is natural to be skeptical about their usefulness in making valid inferences. Examination of the sampling distribution of the eigenvalues and ordered variances from these matrices, however, makes a strong case for them containing much information about underlying associations among genes.

The left panel of Fig. 5(a) is a plot of the positive eigenvalues of the between-epoch covariance matrix. Depicted also are expectations of these eigenvalues had sampling (with the same sample sizes) been done using an identity covariance matrix scaled to have the same trace. The right panel gives a magnified view, showing that the 5th and 95th percentiles lie quite close to the mean, relative to the scale of the observed eigenvalues. Figure 5(b) illustrates similar behavior for the eigenvalues of the within-epoch covariance matrix.

The very large initial eigenvalues are attributable either to large differences in the observed variances of the genes or to clustering of the genes into underlying factors in a factor model. To check to see if the underlying covariance matrices could be diagonal, for each (between- or within-epoch) covariance matrix, we first ordered the observed variances in decreasing order as $(V_1, \ldots, V_p)$ and then subtracted these from the ordered eigenvalues $(L_1, \ldots, L_p)$. As a measure of excess of eigenvalues over ordered variance, we chose the relative differences $(L_j - V_j)/V_j$, which all should be somewhat close to 1 under independence. To see exactly how close these should be under independent normal sampling, we simulated the distributions of the relative differences using independent normal samples. The left panel of Fig. 5(c) shows that the first few eigenvalues from the between-epoch covariance matrix greatly exceed anything that might have been generated under normal independence; the right panel illustrates the relatively narrow band containing 90% of the sampled eigenvalues. For the within-epoch eigenvalues, Fig. 5(d) shows that the first 50–60 of the 756 positive values lie well above the confidence bounds. As described in the next paragraph, the covariance matrices for the rainbow trout study were derived by a somewhat more complicated process than simple random sampling. Simulating data according to this process, while slow, leads to much the same results as those depicted in Fig. 5, except that the confidence bands are a bit wider. Overall, there is very strong empirical evidence that genes are clustered. Assuming this is the case, we now show how to exploit the information about clustering when trying to identify genes that act in concert.

4.2. Select Genes by SCOOP

For each gene, a data matrix of dimension 739 × 9 was formed as follows: each observation at any time in any organ was extended into a row of nine values by fitting a B-spline to that observation supplemented by mean expressions at other time points; the nine values are the coefficients of the fitted B-spline basis functions. Then groups were defined corresponding to the nine basis functions, representing different epochs. Following the SCOOP algorithm, genes were ordered according to the discriminant information. Plotting the discriminant information clearly resulted in an exponential curve (see Fig. 6). The procedure was rerun with the eigenvectors corresponding to the smallest eigenvalues dropped, retaining enough to explain at least 80% of the original variation in the covariance matrix. This produced a different ordering; genes were selected up to the point when the two orderings started to differ. This procedure selected 193 genes.

The swiftness of our algorithm merits some comment. Calculating discriminant information and ranking 16 000 genes occurred within a minute using R on a 2 Gb, 2.3 Ghz machine. Three techniques were especially effective in reducing computational time: first, whenever possible, matrix calculations were used instead of inefficient loops in the R package; second, as discussed in a previous part, singular value decomposition was carried out instead of PCA; third, SCOOP differs from the method of shrunken centroids [1] in the important sense that centroids are not successively shrunk for each gene. Although successive shrinking is very good for the problem of classification by shrunken centroids, it is not at all good in this adaptation for gene selection because information is lost in successive steps leading to too much cumulative distortion. The fortunate consequence is that SCOOP is computationally very fast.

4.3. Clustering of the SCOOP-selected Genes

In the end, 193 genes were selected by SCOOP as demonstrating large, mostly common changes over time in all three organs. Clustering of genes was based on the 27-dimensional (nine sets of rising times [λ_i] for each of three organs) variable importance measures. Agglomerative clustering suggested that the genes be divided into four distinct groups (see Supporting Information Table 1). As a check, K-means clustering was also used to split the genes into four categories, similarly producing two large (67 and 83 gene) clusters and two small (16 and 27 gene) clusters. For each cluster and each organ, a smoothed curve was fitted to the mean expressions over time. Figure 7 gives the mean expression curve for each cluster in the pituitary, ovary, and liver, respectively. The mean time courses have distinct forms in each organ, with the clearest patterns showing up in the pituitary and ovary.

4.4. Related Functions of Genes Selected by SCOOP

By relating the functions of the SCOOP-selected genes, we were able to identify potential pathways, shown in

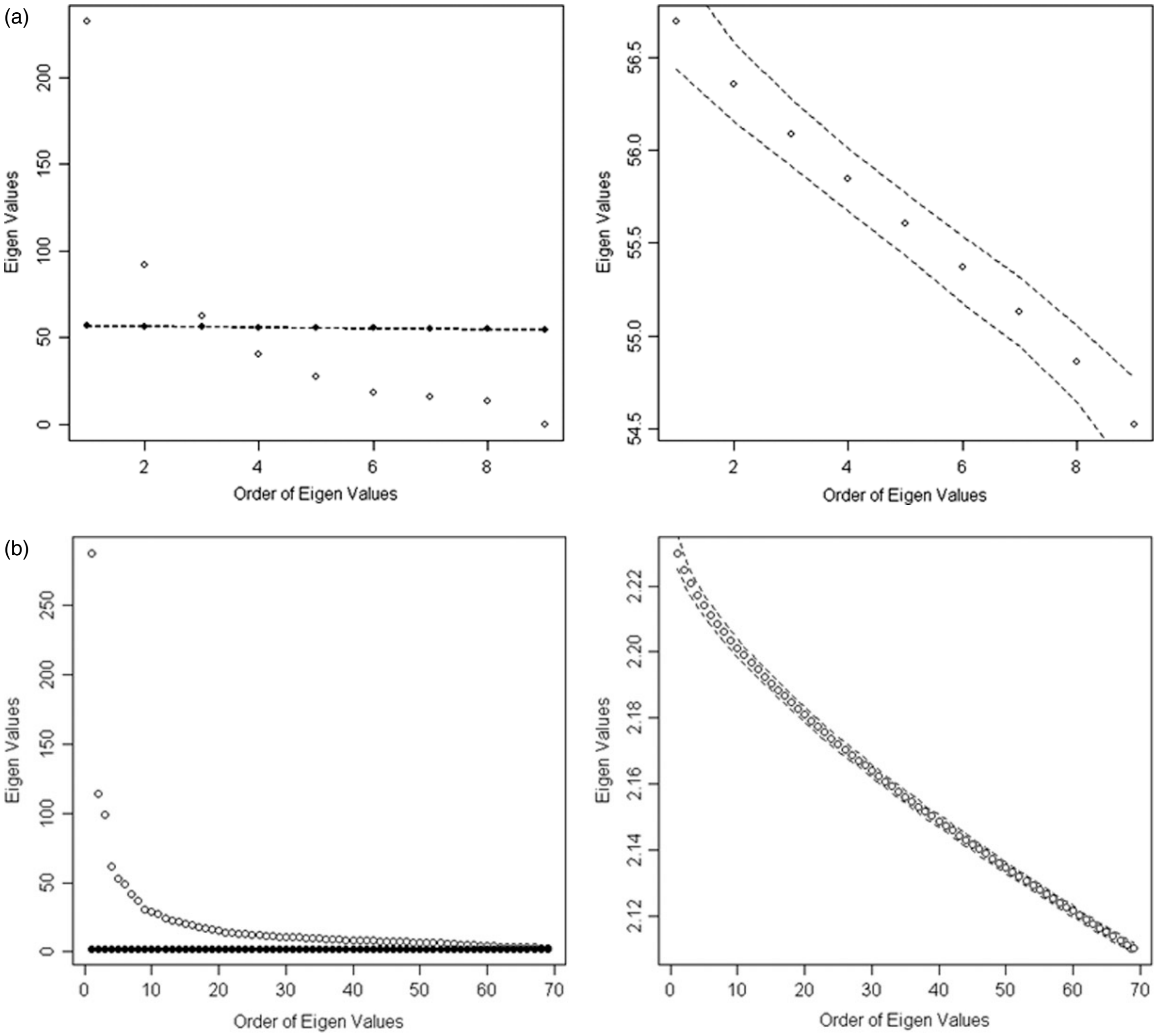

Fig. 5 (a) Observed and simulated eigenvalues of the between-epoch matrix. (b) Observed and simulated eigenvalues of the within-epoch matrix. (c) Simulation of between-epoch relative differences between eigenvalues and ordered variances under independence.(d) Simulation of within-epoch relative differences between eigenvalues and ordered variances under independence.

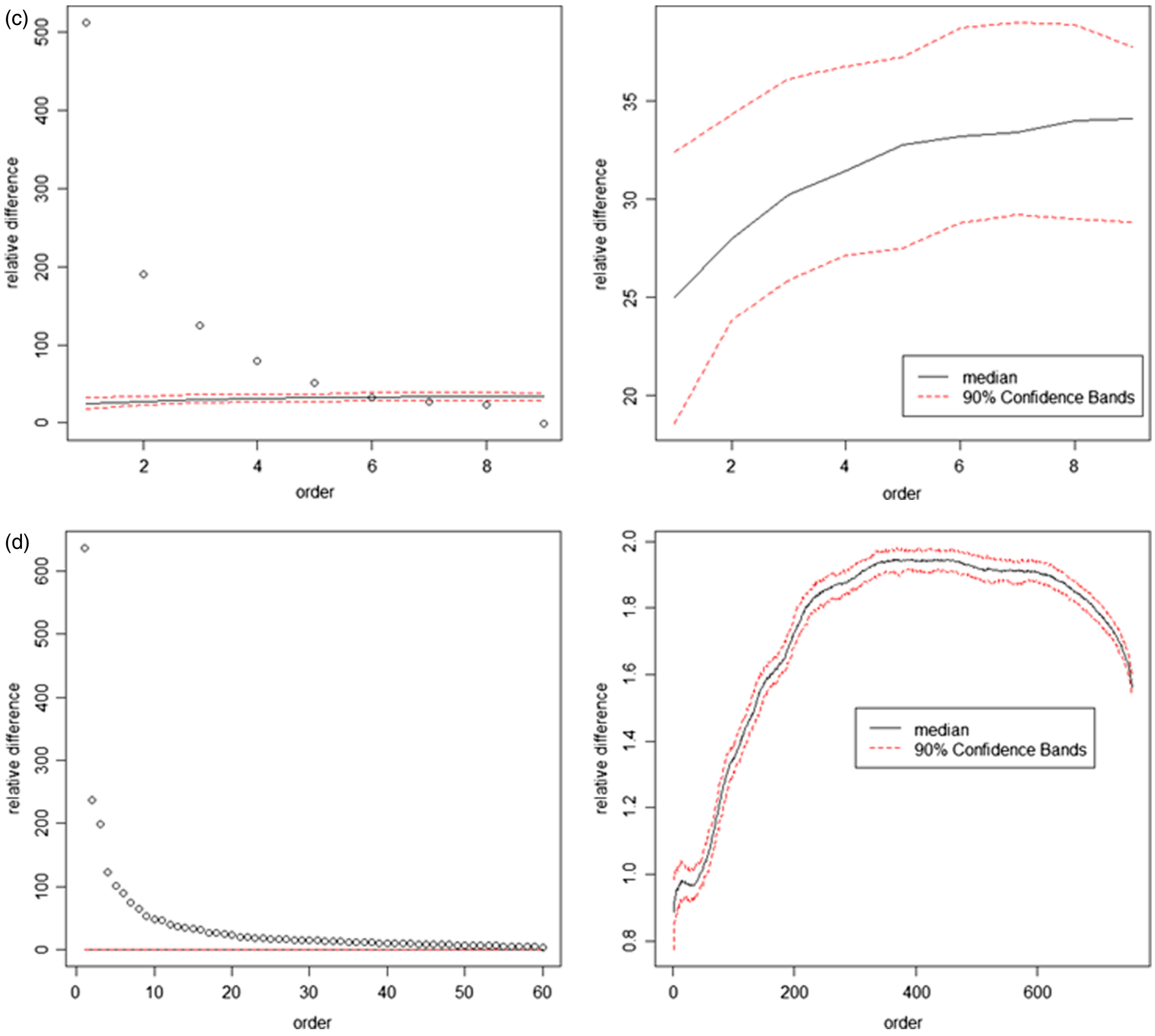

Fig. 5 Continued.

Fig. 8, that support the HPGL signaling network depicted in Fig. 1. Figure 8(a) illustrates pathways starting from gene transcriptional level regulation, RNA synthesis involvement, protein translation and posttranslational modification. The newly translated proteins from these pathways may be categorized by their function: immunology, muscle contraction, reproduction, protein transportation and secretion, metabolism, redox potential regulation, amino acid synthesis and degradation, and neural regulation (Fig. 8(b)). A protein degradation pathway was developed as indicated in Fig. 8(c). A detailed description is available in the supporting information to this paper.

All these pathways are constructed based on biochemical principles [17]. And the idea of how to construct the pathways is in the flow chart of Fig. 8(d). The fact that selected genes do form natural pathways demonstrates that SCOOP can lead to the discovery of an elevated pathway instead of just a few enhanced genes. One noticeable feature of the SCOOP-selected genes is that they reveal several networks, such as immune functioning and metabolic activity, which support the HPGL hormone pathway in addition to the direct pathway itself. Immune system involvement has also been suggested in several studies of environmental exposure to estradiol [18,19]. An elevated metabolic activity is required to provide enough energy for the reproduction cycle. These previously overlooked genes should draw attention in future studies.

4.5. Comparing EDGE-selected Genes and SCOOP-selected Genes

On the basis of the EDGE analysis, a set of 55 ovarian genes and a set of 67 hepatic genes were selected as having significant changes over time (see Supporting Information).

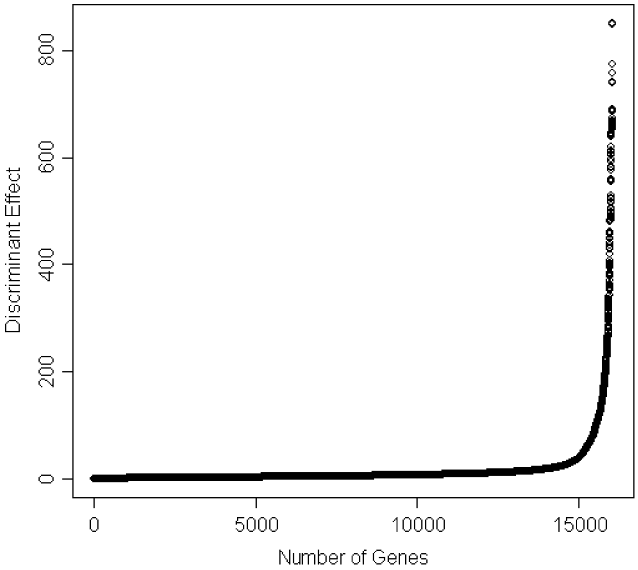

Fig. 6 The plot of discriminant effect versus the rank of genes by the Shrunken Centroid Ordering by Orthogonal Projections method.

However, EDGE did not detect any genes in the pituitary that were differentially expressed over time. When expression values were combined over all three organs to identify common differentially expressed genes, as had been done with SCOOP, no genes were discovered by EDGE.

5. DISCUSSION

Using the new method SCOOP, we identified related genes that are differentially expressed over time in each of three different organs that are part of the HPGL axis in the female rainbow trout. The time course pattern in all three organs reveals how each cluster distinctly relates to this process. In the pituitary, all selected genes undergo dramatic changes at about 130 days into the reproductive cycle, but cluster 4 was up-regulated whereas the other three clusters were downregulated. This is consistent with the increased concentrations of hormones like FSH in the blood of females [20], for example, and thus the identified genes appear highly correlated with the reproductive process. That the earliest change of gene activity occurs in the pituitary makes reproductive sense because this is the organ that produces hormones (i.e., FSH) that initiate ovarian development. In the ovary, the genes identified showed peak activity around 280 days which conforms to the height of ovarian growth due to oocyte uptake of vitellogenin from the liver. Additionally, the initiation of the expression of genes involved in final oocyte maturation would be expected too (e.g., 17,20β-dihydroxy- 4-pregene-3-one) [9,21,22]. This further implicates the selected genes in the reproductive process. Interestingly at around cycle day 250, the cluster 4 switched from being out of phase with the other clusters to being in the same high expression state around cycle day 290. The patterns of the clusters in liver suggest a general dampening of expression over time, perhaps suggesting less involvement of these genes in other processes as spawning time nears. With the multifunctional role of the liver in many physiological processes parsing out distinct reproductive involvement appears difficult. Despite this, the general expression patterns in both ovary and pituitary, and to a lesser extent in the liver, suggest direct or indirect reproductive significance for SCOOP-selected genes. Unlike other methods, SCOOP critically uses information about the correlation of expression among genes. Thus, it can identify potential pathways even without annotation, and thereby provide a self-contained "gene enrichment." This feature is especially useful in the present context of GRASP arrays, which provide annotation for only about half of the genes. In conjunction with Friedman's new GPS method, here used to cluster genes robustly, SCOOP may offer a new way to characterize genes in different biological process and further advance this field.

We also used EDGE analysis to discover genes common to the three organs. The reason behind EDGE's failure to detect the linkage between organs is its focus on each gene separately. Although this may be appropriate for small sample sizes for which correlations are not well estimated, overlooking the correlation among genes in the current setting would have missed most of the processes that support the reproductive cycle. Moreover, results based on the F-test make the EDGE method sensitive to the distribution of the data. In many cases, microarray data may not be distributed normally. In contrast, SCOOP does not require normally distributed data. By preserving the correlation among genes, it remains reliable under any conditions for which the correlation matrix is well estimated. Most importantly, EDGE tends to discover genes instead of pathways, making the results harder to interpret. Moreover, it tends to have a high false negative rate. Here it failed to identify genes like the FSH beta subunit which is an important reproductive gene [23]. Nevertheless, EDGE produces basically sound findings. Despite its selection of a different set of genes than those selected by SCOOP, several of the EDGE-selected genes may also be fitted to the pathways we constructed in Fig. 8, for example, immunology, signal transduction, metabolism, protein degradation etc. Although these appear less clear for EDGE than for SCOOP, they serve as further confirmation of the pathways inferred from SCOOP. Also, EDGE identified genes like the *O. mykiss* vtg1 (i.e., vitellogenin) gene, filling the gap between linkage of liver and other organs in mathematical models [15,24] of the HPGL signaling network (Fig. 1), which was missed in SCOOP.

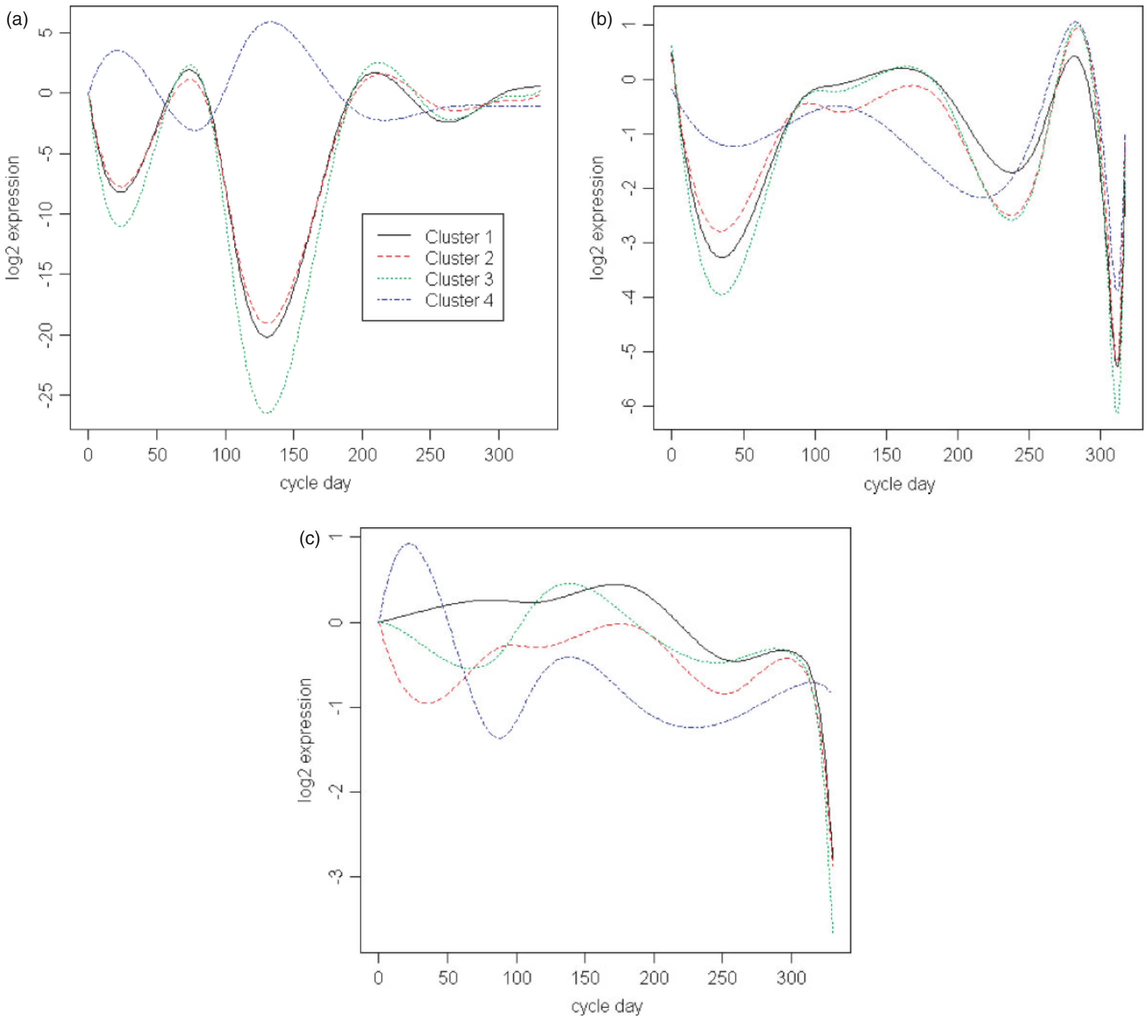

Fig. 7 (a) The mean expression curves of each of the Shrunken Centroid Ordering by Orthogonal Projections selected gene clusters in the female rainbow trout pituitary over the reproductive cycle. (b) The mean expression curves of each of the Shrunken Centroid Ordering by Orthogonal Projections selected gene clusters in the female rainbow trout ovary over the reproductive cycle. (c) The mean expression curves of each of the Shrunken Centroid Ordering by Orthogonal Projections selected gene clusters in the female rainbow trout liver over the reproductive cycle.

Thus, EDGE and SCOOP appear to have complementary virtues.

Despite its usefulness, SCOOP did not identify genes such as LH beta subunit in the pituitary, which is known to be important in the female rainbow trout reproductive process [25]. There are several possible explanations: First, regulation may happen at the translational level instead of at the transcriptional level, since this hormone is made up of two different protein subunits, α and β. This means that the mRNA expression level of one of the subunits may not reflect the expression of the complete translated LH protein [26]. Second, the high noise nature of the microarray data—SCOOP might be improved by estimating the principal eigenvectors of the covariance matrices more robustly. Potential remedies involve using microarray platforms that are more comprehensive, containing both subunit mRNAs for hormone targets that are complex proteins, for example, and/or that utilize significant probe redundancy (i.e., multiple microarray probes/gene) to increase precision.

Although the association between clusters and pathways was not perfect, the clustering was suggestive enough to construct some pathways that provide global insight into the molecular biology of the reproductive system of rainbow trout. The K-means clusters, indicated by the four different colors in Fig. 7(a– c), are strongly related to the identified pathways. As seen from Fig. 8(a), cluster 1 is involved in transcriptional regulation at the DNA level, and encodes some ribosomal subunits. In contrast, cluster 3 is involved in regulation at the RNA level and some posttranslational modification. Cluster 2 has only a minor effect here, but

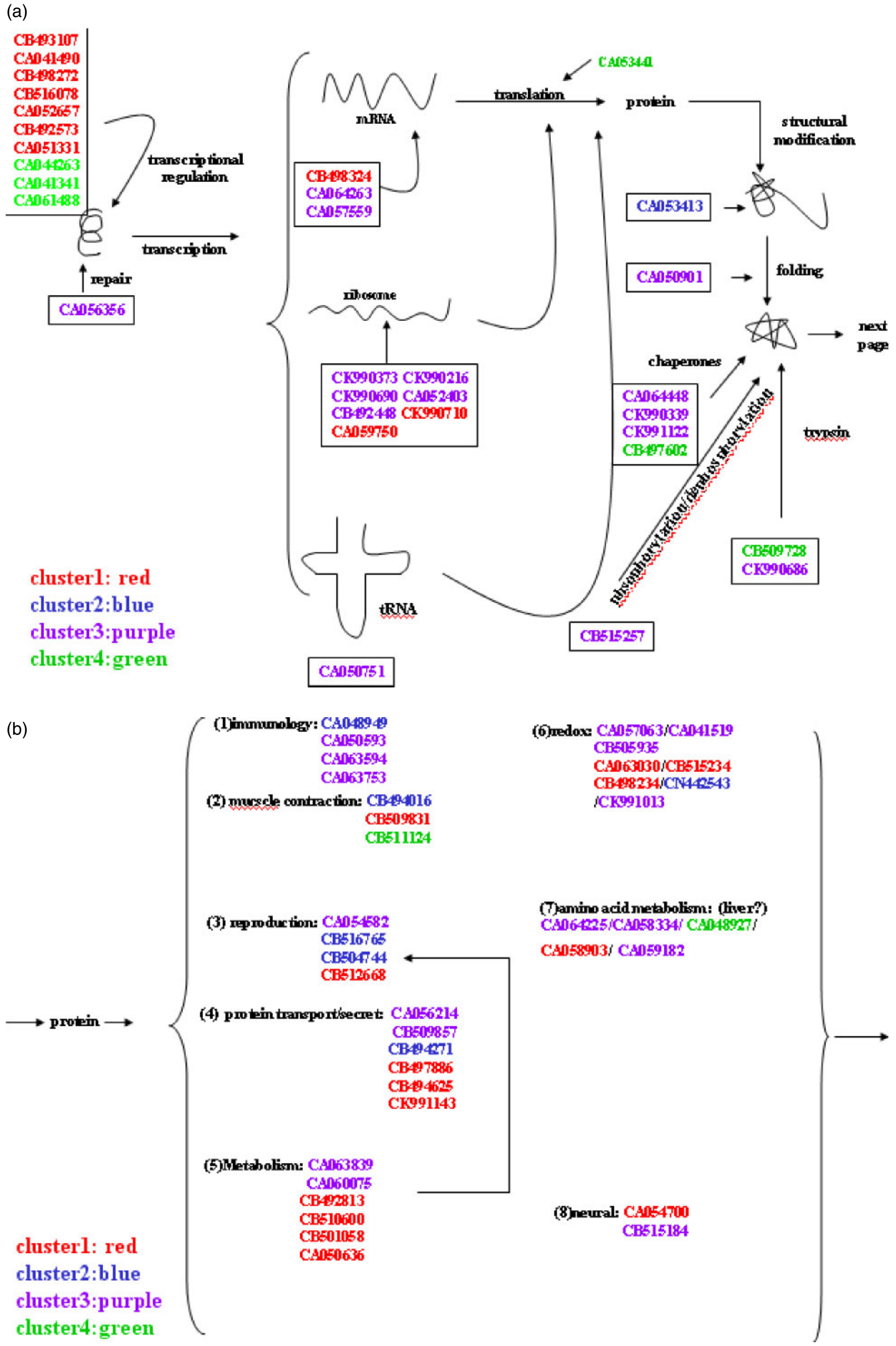

Fig. 8 (a) Some common genes selected by the Shrunken Centroid Ordering by Orthogonal Projections method clustering in the female rainbow trout hypothalamus–pituitary–gonad–liver signaling network functionally organized according to a transcription, translation, and post transcriptional modification pathway. (b) Some common genes selected by the Shrunken Centroid Ordering by Orthogonal Projections method clustering in the female rainbow trout hypothalamus–pituitary–gonad–liver signaling network listed based on protein function. (c) Some common genes selected by the Shrunken Centroid Ordering by Orthogonal Projections method clustering in the female rainbow trout hypothalamus–pituitary–gonad–liver signaling network placed in a protein degradation pathway. (d) The central idea behind the pathway construction.

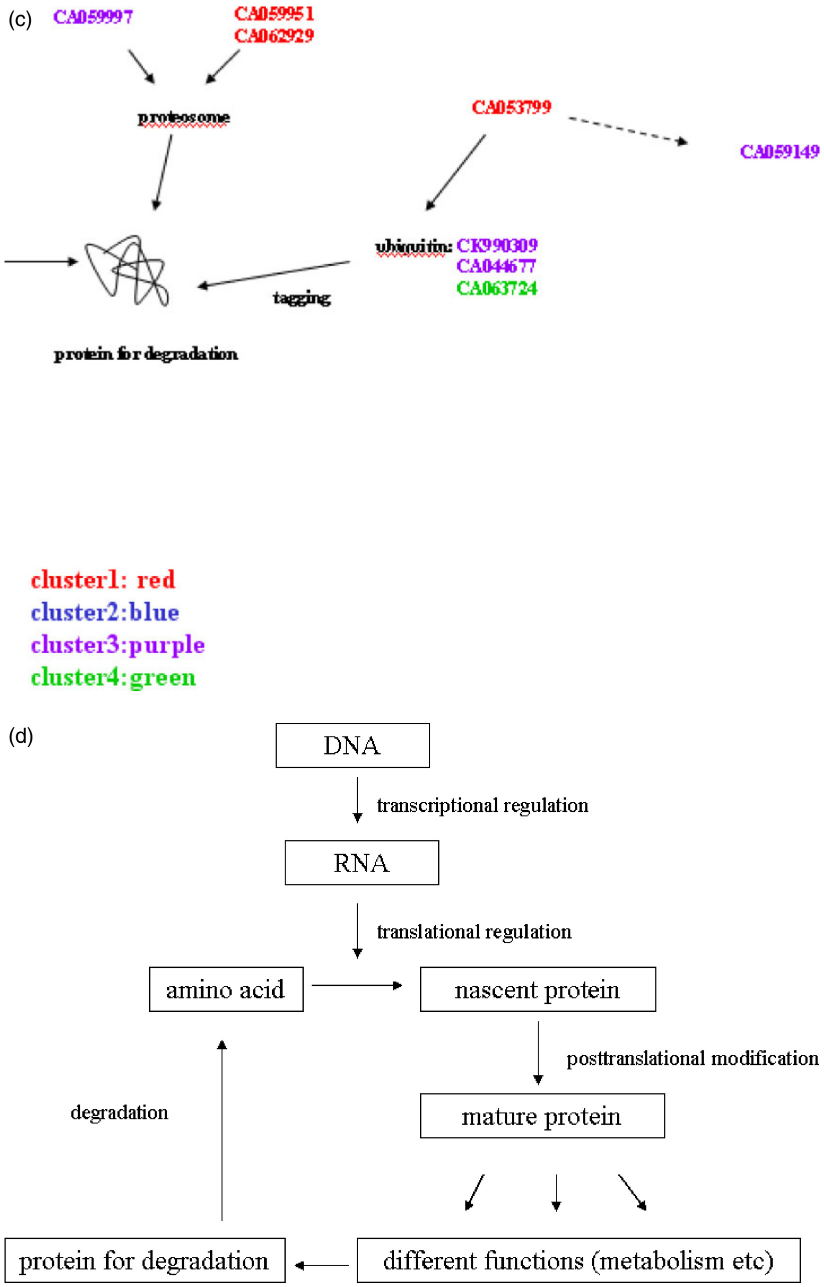

Fig. 8 Continued.

cluster 4 contains the remaining transcriptional factors. However, clusters do not separate the pathways as clearly in Fig. 8(b) because these types of functions usually require genes to be expressed sequentially in a cascade. Despite this, most immunology genes come from the cluster 3 family. For metabolism genes, most come from the cluster 1 family. Redox regulatory genes mostly split between clusters 1 and 3. Similarly, Fig. 8(c) shows that clusters 1 and 3 are also essential for regulation at the translational level (i.e. protein degradation). Thus, cluster 1 is essential for regulation at both transcriptional and translational levels while cluster 3 is vital only for the translational level. Cluster 4 is crucial for transcriptional regulation. Cluster 2 may be important for performing various functions instead of regulation.

In Fig. 8(b), the SCOOP-selected genes were grouped according to each individual function. However, the regulation of these functions is specific to each organ. That is, functional groups are up-regulated and downregulated at different times in each organ. The unique time course patterns of gene activity in each organ provide new insight into the mechanisms controlling reproduction in rainbow trout.

The temporal clustering of genes related to the reproductive cycle presented in this analysis may be useful in expanding the mathematical model of the HPGL axis [27]. The current mathematical model focuses on the hormonal regulation of final oocyte maturation and spawning, but does not model changes in gene expression outside of a few key hormonal components. This model also does not incorporate the liver, which is an important site for hormone metabolism and vitellogenesis (see Fig. 1). Candidate genes identified in this analysis could be tested in the mathematical model as additional variables to give a better understanding of the biological mechanisms that lead to the observed changes in gene expression over time. Including these candidate genes in the model could guide the formation of testable biological hypotheses as well as allow incorporation of future experimental data. Modeling may reveal possible mechanisms by which some genes are up-regulated at a particular time while others are downregulated. For example, genes in one cluster might be under the direct control of gonadotropin-releasing hormone (GnRH), whereas another cluster may be regulated more indirectly through complex feedback loops. Ultimately all these suggested pathways and models need to be verified experimentally, but the analytical tools discussed provide guidance, and may save much time, effort and resource in discovering new molecular networks.

5.1. Acknowledgements

We thank the National Science Foundation for support under grant DMS-0540693 and the referees for their constructive comments.

REFERENCES

[1] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, Significance analysis of time course microarray experiments, Proc Natl Acad Sci 102 (2005), 12837–12842.

[2] D. Baron and Y. Guiguen, Gene expression during gonadal sex differentiation in rainbow trout (*Oncorhynchus mykiss*): from candidate gene studies to high throughput genomic approach, Fish Physiol Biochem 28 (2004), 119–123.

[3] J. A. Luckenback, D. B. Iliev, F. W. Goetz, and P. Swanson, Identification of differentially expressed ovarian genes during primary and early secondary oocyte growth in cohosalmon (*Oncorhynchus kisutch*), Reprod Biol Endocrinol 6 (2008), 2.

[4] J. T. Popesku, C. J. Martyniuk, J. Mennigen, H. Xiong, D. Zhang, X. Xia, A. R. Cossins, and V. L. Trudeau, The goldfish (*Carassius auratus*) as amodel for neuroendocrine signaling, Mol Cell Endocrinol 293 (2008), 43–56.

[5] N. Bromage and R. Cumaranatunga, Egg production in the rainbow trout, In Recent Advances in Aquaculture, J. F. Muir, R. J. Roberts, and C. Helm, eds. London, Timber Press, 1988, 63–138.

[6] N. M. Sherwood, D. B. Parker, J. E. Mcrory, and D. W. Lescheid, Molecular evolution of growth hormone-releasing hormone and gonadotropin-releasing hormone, In Fish Physiology Molecular Endocrinology of Fish, N. M. Sherwood and C. L. Hew, eds. San Diego, Academic Press, 1994, 3–66.

[7] P. Swanson, J. T. Dickey, and B. Campbell, Biochemistry and physiology of fish gonadotropins, Fish Physiol Biochem 28 (2003), 53–59.

[8] G. Young, M. Kusakabe, I. Nakamura, P. M. Lokman, and F. W. Goetz, Gonadal steroidogenesis in teleost fish, In Hormones and Their Receptors in Fish Reproduction, Vol. 2, N. Sherwood and P. Melamed, eds. Singapore, World Scientific Press, 2005, 155–223.

[9] R. A. Wallace, Vitellogenesis and oocyte growth in nonmammalian vertebrates, In Developmental Biology, L. W. Browder, ed. New York, Plenum Press, 1985, 127–177.

[10] Z. Bar-Joseph, G. Gerber, and D. K. Gifford, Continuous representations of time series gene expression data, J Comput Biol 10 (2003), 341–356.

[11] Y. C. Tai and T. P. Speed, On the gene ranking of replicated microarray time course data, http://www.stat.berkeley.edu/tech-reports/735.pdf, 2007.

[12] Y. C. Tai and T. P. Speed, A multivariate empirical Bayes statistic for replicated microarray time course data, Ann Stat 34 (2006), 2387–2412.

[13] M. J. Nueda, A. Conesa, J. A. Westerhuis, H. C. Hoefsloot, A. K. Smilde, M. Talón, and A. Ferrer, Discovering gene expression ptterns in time course microarray experiments by ANOVA-SCA, Bioinformatics 23 (2007), 1792–1800.

[14] D. Telesca, L. Y. T. Inoue, M. Niera, R. Etzioni, M. Gleave, and C. Nelson, Differential expression and network inferences through functional data modeling, Biometrics (2009) (in press). 2008/DOI: 10.1111.

[15] S. E. Hook, A. D. Skillman, J. A. Small, and I. R. Schultz, Gene expression patterns in rainbow trout, *Oncorhynchus mykiss*, exposed to a suite of model toxicants, Aquat Toxicol 77 (2006), 372–385.

[16] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed, Normalization for cDNA microarray data, SPIE BiOS, San Jose, CA, Technical Report, January, 2001.

[17] A. Lehninger, D. L. Nelson, and M. M. Cox, Lehninger Principles of Biochemistry (4th ed.), New York, Macmillan, 2005, 116–786.

[18] C. R. Wira and D. A. Sullivan, Effect of estradiol and progesterone on the secretory immune system in the female genital tract, Adv Exp Med Biol 138 (1981), 99–111.

[19] M. J. Myers, L. D. Butler, and B. H. Petersen, Estradiol-induced alteration in the immune system. II. Suppression of cellular immunity in the rat is not the result of direct estrogenic action, Immunopharmacology 11 (1986), 47–55.

[20] F. Prat, J. P. Sumpter, and C. R. Tyler, Validation of radioimmunoassays for two salmon gonadotropin (GTH I and GTH II) and their plasma concentrations throughout the reproductive cycle in male and female rainbow trout, Biol Reprod 54 (1996), 1375–1382.

[21] J. V. Planas, J. Athos, F. W. Goetz, and P. Swanson, Regulation of ovarian steroidogenesis in vitro by follicle simulating hormone and luteinizing hormone duirng sexual maturation in Salmonid fish, Biol Reprod 62 (2000), 1262–1269.

[22] M. K. Koldras, D. Bieniarz, and E. Kime, Sperm production and steroidogenesis in testes of the common carp, *Cyprinus*

carpio L., at different stages of maturation, J Fish Biol 37 (1990), 635–645.

[23] J. T. Dickey and P. Swanson, Effect of sex steroid on gonadotropin (FSH and LH) regulation in coho salmon, J Mol Endocrinol 21 (1998), 291–306.

[24] V. Trichet, N. Buisine, N. Mouchel, P. Morán, A. M. Pendás, J.-P. LePennec, and J. Wolff, Genomic analysis of the vitellogenin locus in rainbow trout (*Oncorhynchus mykiss*) reveals a complex history of gene amplification and retroposon activity, Mol Gen Genet 263 (2000), 828–837.

[25] D. L. Villeneuve, P. Larkin, I. Knoeb, A. L. Miracle, M. D. Kahl, K. M. Jensen, E. A. Makynen, E. J. Durhan, B. J. Carter, N. D. Denslow, and G. T. Ankley, A graphical systems model to facilitate hypothesis-driven ecotoxicogenomics research on the teleost brain–pituitary–gonadal axis, Environ Sci Technol 41 (2007), 321–330.

[26] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc Natl Acad Sci 97 (2000), 262–267.

[27] J. Kim, W. L. Hayton, and I. R. Schultz, Modeling the brain–pituitary–gonad axis in salmon, Mar Environ Res 62 (2006), S426–S432.