Inferring metabolic pathway activity levels from RNA-Seq data

Yvette Temate-Tiagueu*[1], Sahar Al Seesi[3], Meril Mathew[2], Igor Mandric[1], Alex Rodriguez[2], Kayla Bean[2], Qiong Cheng[4], Olga Glebova[1], Ion Măndoiu[3], Nicole B. Lopanik[2],[5] and Alexander Zelikovsky*[1]

[1]Department of Computer Science, Georgia State University, 34 Peachtree St., 30303, Atlanta, GA, USA [2]Department of Biology, Georgia State University, 100 Piedmont Ave., 30303, Atlanta, GA, USA [3]Computer Science & Engineering Department, University of Connecticut, 06269, Storrs, CT, USA

[4]Department of Pharmacology, University of Miami, Miami FL, USA [5]Current address: School of Earth and Atmospheric Sciences, School of Biology, Georgia Institute of Technology, 311 Ferst Dr., 30332, Atlanta, GA, USA

Email: Yvette Temate-Tiagueu - ytematetiagueu1@cs.gsu.edu; Sahar Al Seesi - sahar@engr.uconn.edu; Meril Mathew - mmathew3@student.gsu.edu; Igor Mandric - imandric1@student.gsu.edu; Alex Rodriguez - arodriguez27@student.gsu.edu; Kayla Bean - kbean1@student.gsu.edu; Qiong Cheng - qxc@miami.edu; Olga Glebova - oglebova1@student.gsu.edu; Ion Măndoiu - ion@engr.uconn.edu; Nicole B. Lopanik - nicole.lopanik@eas.gatech.edu; Alex Zelikovsky - alexz@cs.gsu.edu;

*Corresponding author

Abstract

Background

Assessing pathway activity levels is a plausible way to quantify metabolic differences between various conditions. This is usually inferred from microarray expression data. Wide availability of NGS technology has triggered a demand for bioinformatics tools capable of analyzing pathway activity directly from RNA-Seq data. In this paper we introduce XPathway, a set of tools that compares pathway activity analyzing mapping of contigs assembled from RNA-Seq reads to KEGG pathways. The XPathway analysis of pathway activity is based on expectation maximization and topological properties of pathway graphs.

Results

XPathway tools have been applied to RNA-Seq data from the marine bryozoan *Bugula neritina* with and without its symbiotic bacterium "*Candidatus* Endobugula sertula". We successfully identified several metabolic pathways with differential activity levels. The expressions of enzymes from the identified pathways have been further validated through quantitative PCR (qPCR).

Conclusions

Our results show that XPathway is able to detect and quantify the metabolic difference in two samples. The software is implemented in C, Python and shell scripting and is capable of running on Linux/Unix platforms. The source code and installation instructions are available at http://alan.cs.gsu.edu/NGS/?q=content/xpathway.

Introduction

For the past several years, RNA-Seq has revolutionized biological research through the many advantages it provides. Because of RNA-Seq, it is easier to characterize transcripts and their isoforms, to detect genes without need of prior information in the form of probes or primers, and estimate expression level of transcript with good precision.

In contrast to microarray data, RNA-Seq data allows frequency of expression of all transcripts without a priori knowledge of the gene sequence. RNA-Seq data can account for the entire RNA volume producing enzyme for a given pathway. When applied to metatranscriptome data, the first challenge of pathway analysis is to decide which metabolic pathways are active in the sampled community (i.e., pathway activity detection). Recent software tools (*MEGAN4* [1] and *MetaPathways* [2] using SEED and KEGG (Kyoto Encyclopedia of Genes and Genomes) [3] annotations) enable the organization of transcripts or reads into ortholog groups and pathways by collecting all transcripts or reads represented by at least one ortholog group and providing that collection to the user. The parsimonious approach *MinPath* [4] identifies the smallest family of pathways covering all expressed ortholog groups. A more elaborate Markov Chain Monte Carlo (MCMC) approach takes into account the co-occurrences of genes in more than one pathway for analyzing metagenomic data [5]. Following pathway detection, the second major challenge of pathway analysis is to infer pathway activity levels to enable detection of differential expression. Few existing tools incorporate this step, a major focus of this paper.

Methods that treat pathways as simple gene sets [6, 7] are popular even though they do not use all information available. In recent years, a number of pathway analysis methods have been developed that combine knowledge of pathway topology (e.g., gene position on the pathway, gene-gene interactions, etc.) with gene expression data based on comparative analyses (reviewed in [8]). Such methods have been applied primarily to experimental studies of single organisms. Despite the inherent pathway architecture of microbial biochemical function, relatively few analyses of complex metatranscriptomic datasets incorporate pathway-level inference of metabolic activity.

The new analysis techniques, presented here, is suitable for RNA-Seq data analysis to investigate the underlying metabolic differences between species living in separate environments.

Our contribution consists of the following:

1. A novel graph-based approach to analyze pathways significance. We represent metabolic pathways as graphs that use nodes to represent biochemical compounds, with enzymes associated with edges describing biochemical reactions.

2. An implementation of an EM algorithm, in which pathways are viewed as sets of orthologs.

3. The validation of the two approaches through differential expression analysis at the transcripts and genes levels and also through real-time quantitative PCR experiments.

Pathways can also be regarded as a set of ortholog group on which we can apply a set cover. We will use a binary ortholog group expression model to determine if there is or not RNA-Seq evidence for the expression of a given ortholog group in a given sample.

The validation step of these methods consists of extracting the genes involved in our estimated differential pathways activity levels, and analyzing their expression levels or transcript frequency estimation. We expect to see the differential pathway activity confirmed at the protein and contigs level. We carry this final analysis through the novel bootstrapping tool IsoDE [9].

Our experimental study was performed with RNA-Seq data from the marine bryozoan, *Bugula neritina*. Using the two novel computational approaches we implemented, we were able to find differentially expressed pathways from the data. This result is been validated by quantitative PCR (qPCR) conducted using a housekeeping gene also identified in the data. The housekeeping gene, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), was chosen to normalize the qPCR data, as is standard practice. Based on our results, we applied qPCR experiments to quantify transcripts in the following identified pathways: Fatty acid elongation (ko00062), Peroxisome (ko04146) and Ribosome biogenesis in eukaryotes (ko03008) [3].

The rest of the paper is organized as follows. Our formal models to analyze pathways and to infer pathway activity are presented in the next section entitled Methods. Following Methods is the Differential-analysis section in which we present how we compute differential activity between pathways. We finish by presenting and discussing our results on *Bugula neritina* data in the Results section. The paper is concluded with possible future work.

Methods

In this paper we introduce a graph-based and an expectation maximization approach to identify specific differences between biological systems on the level of ortholog groups and pathways.

Figure 1 presents the entire flow of XPathway tools. In the graph-based approach, we compute a p-value using parameters extracted from the network to answer two different statistical questions: (1) When and based on what parameter can we say that a set of proteins significantly map to a pathway? (2) What is the probability of finding such a mapping by chance given the data (transcripts/reads/proteins) and a pathway topology? Finally, significant metabolic pathways are selected by comparing the p-value of the original pathway with the ones from different bootstrapped samples. The expectation maximization method on the other hand uses the interaction among identified ortholog groups to decide on a pathway activity. The last part of the flow consists of validating both branches. First, we compute differential expression analysis on all contigs extracted from pathways output by both branches. Secondly, a qPCR experiment is carried out on the contigs which have a fold change of 1.5 or more.

Expectation maximization model of pathway activity

In this section we present an EM-based algorithms for inferring pathway activity levels based on metatranscriptome sequence data. Let w be a pathway that is considered to be a set of enzymes represented by their ortholog groups $w = \{p_1, \ldots, p_k\}$. Since an ortholog group can have multiple functions and participate in multiple pathways, the pathways can be viewed as a family of subsets W of the set of all ortholog groups P. Below we start by introducing a uniform binary pathway activity model based on a discrete ortholog group expression model.

The uniform binary pathway activity model is based on the assumptions of *uniformity*, namely that each molecule from an ortholog group participates in each active pathway with the same probability (i.e., in equal proportions) and of *binary activity*, which postulates that a pathway is active if the level of ortholog group activity exceeds a certain (possibly pathway dependent) threshold. Formally, let $\delta(w)$ be a binary variable indicating the *activity status* of w, i.e., $\delta(w) = 1$ if w is active and $\delta(w) = 0$, otherwise. Also, let the *activity level* of pathway w be the summation over constituent ortholog groups g of their participation g_w in w. Since we assume that each ortholog group g is equally likely to participate in each pathway containing it, it follows that $g_w = (1 + \sum_{w' \ni p, w' \neq w} \delta(w'))^{-1}$ and the activity level f_w of pathway w is given by

$$f_w = \sum_{g \in w} g_w = \sum_{g \in w} \frac{1}{1 + \sum_{w' \ni g, w' \neq w} \delta(w')} \tag{1}$$

The binary activity status of w is computed from its activity level f_w and the threshold T_w as follows

$$\delta(w) = \begin{cases} 0 & \text{if } f_w < T_w \\ 1 & \text{if } f_w \geq T_w \end{cases} \tag{2}$$

The uniform binary model described by equations (1)-(2) can be solved using a simple iterative algorithm. The algorithm starts with assigning activity status $\delta(w) = 1$ to each pathway $w \in W$, i.e., $\Delta^0(W) = \{\delta^0(w) | w \in W\} \leftarrow 1$ and then repeatedly updates the activity level according to (1) and the activity status according to (2). The procedure terminates when the status sequence $\Delta^0(W) = 1, \Delta^1(W), \Delta^2(W), \dots$ starts to oscillate $\Delta^{n+k}(W) = \Delta^n(W)$ or converges. In all our preliminary experiments, an oscillation with period $k = 2$ is achieved in at most 10 iterations. Also the threshold T_w does not significantly change the order of pathways sorted with respect to their activity levels estimated as the mean f_w after convergence. The model is represented in Figure 2.

Although the uniform binary model allows the computation of pathway activity by assigning ortholog groups to pathways, it does have some limitations hindering it for capturing specific attributes of the metabolic network. For example, the binary uniform model assigns only value 1 or 0, if the ortholog group belongs to a pathway or not, respectively. This yes or no assumption is not always true since there may be a fractional part of an ortholog group belonging to different pathways. Moreover, the uniformity model is not easily applicable to natural processes because all assignment are never equally likely. Finally the model is not completely stable but rather periodic with some subsets of ortholog groups fluctuating between pathways.

Graph-based estimation of pathway significance

Ideally, a comprehensive pathway analysis method would take into consideration the position and role of each gene in a pathway, the efficiency with which a certain reaction is carried out, and some limiting factors (e.g. dealing with metagenomics data or not). With genome data, it is possible to consider pathways size, gene length and overlap in gene content among pathways [5] to compute the relative abundance of pathways and pathway ranking, but this approach might not work with RNA-Seq data especially in the absence of a genome reference.

Henceforth, in our second approach, each pathway is viewed as a network of enzymes also called EC numbers (Enzyme Commission numbers) in order to compute their statistical significance. Significance of a pathway activity on the sample is measured by the randomness of the positions of matched enzymes in the corresponding KEGG pathway graph. The randomness is measured using a permutation model for finding significant pathway alignments and motifs [10].

This model assumes that the subset of expressed enzymes in an active annotated pathway should be connected. The enzyme permutation model finds the average vertex degree in the subgraph induced by expressed enzymes. Then the same parameter is computed for sufficiently many random permutations of enzyme labels. The statistically significant match should have density higher than 95% of permutations. Specific characteristics of the graph taken into account in our analysis are:

1. Number of nodes. A node represents a protein that got mapped during BLAST. KEGG usually assigns a green color to those proteins in the graph.

2. Density = (Number of edges) / (Number of vertex - 1)

3. Number of 0 in and out-degree vertex. Let call this number x. x is defined by:

   x = ((number of vertex with out-degree = 0) + (number of vertex with in-degree=0)) / 2 * (number of nodes)

4. We also consider other criteria such as (1) Number of green connected components, (2) Largest Number of nodes in connected component and (3) Largest Number of edges in connected component.

Using these metrics, we compute the density of the induced graph composed of only mapped proteins. We obtain the names of those proteins through EC numbers on the graph. Below, we present two graph-based models, the vertex label swapping and the edge swapping for random graph generation, to analyze pathways. This model is explained by the left side of Figure 1 and Figure 3 presents an example when we permute labels of two vertices.

Model 1: Vertex label swapping for random graph generation

In this model, we keep the same topology but we allow swapping of labels between two vertices. One known issue of this approach is, vertices with high degree always get connected. This might lead to too many significant matches, thus increasing the false positive rate. The vertex label swapping algorithm can be represented as follows:

Model 2: Edge swapping for random graph generation

Algorithm 1 Vertex labels swapping algorithm for random graph generation

```
$G = (V, E) \leftarrow$ Original Graph
$i \leftarrow 1$
$j \leftarrow 1$
for $i \leq m$ do
    for $j \leq n$ do
        randomly pick two vertices a and b from V
        if no $(label(a) == label(b))$ then
            swap label of a and b
        end if
    end for
end for
```

Because of the bias in the vertex label swapping model, we also implemented the edge swapping. Here, the idea is to keep the in-degree and out-degree of each node the same, swapping nodes only if these values do not change. We keep vertex labels the same. Figure 4 presents an example when we permute two edges. The edge swapping algorithm can be represented as follows:

Algorithm 2 Edge swapping algorithm for random graph generation

```
$G = (V, E) \leftarrow$ Original Graph
$i \leftarrow 1$
$j \leftarrow 1$
for $i \leq m$ do
    for $j \leq n$ do
        randomly pick two edges a ($a0 \rightarrow$a1) and b($b0 \rightarrow$b1) from E
        if no $(a0 == b0)$ or $(a0 == b1)$ or $(a1 == b0)$ or $(a1 == b1)$ and no ($a1 \rightarrow$a0 or $b1 \rightarrow$b0) then
            remove edges a and b
            create new edges: $a0 \rightarrow$b1 and $b0 \rightarrow$a1
        end if
    end for
end for
```

Differential analysis of pathway activity and significance

Differential analysis of pathway activity

The goal of this analysis is to determine which pathway needs to be considered more closely to understand the difference in the metabolism of two organisms. For this purpose, we use the pathway expression computed from the binary model presented earlier. First we compute expression of each pathway present in the set of pathways we get from KEGG for a given sample. Then we compute the difference between the expression of each pathway. Under this model, the pathways selected as having differential activity are the ones where the ratio of their expression is greater than a certain threshold. Table 3 presents our results on differential analysis of pathway activity.

Differential analysis of pathway significance

Differential analysis of pathway significance is based on the *p-value* described in the graph-based sub-section of Methods. We randomly permute each pathway graph generating m different graphs. Note that even the smallest pathway graphs contains at least 15 nodes and about 40 edges which is sufficient to generate default $m = 200$ distinct random graphs. A pathway is significant if the *p-value* of the mapping is less than 5%. The *p-value* is the position of the original graph when placed in the sorted list of all randomly generated graphs sorted first by "density" (largest to smallest) and then by the number of nodes having 0 in-degree or 0 out-degree (smallest to largest). A pathway is *significant* if its *p-value* is less than 5%, *very significant* if its *p-value* is less than 1% and *the most significant* if its *p-value* is less or equal to 0.5%.

Let $p1$ be the *p-value* for pathway X in sample 1 and let $p2$ be the *p-value* for pathway X in sample 2. We say that pathway X is differentially significant between the two samples if the probability computed by the equation of $probDiff(X)$ below is greater than 50%.

$$probDiff(X) = \begin{cases} (1 - p1) * p2 & \text{if } p1 \geq p2 \\ (1 - p2) * p1 & \text{if } p2 < p1 \end{cases}$$

For example, let us consider $m = 200$ randomly generated graphs and the vertex label swapping model. In Figure 5 representing part of the Fatty acid elongation pathway (ko00062), the mapped enzymes (filled rectangles) in sample 1 form a sub-graph with density = 1.875 and the number of 0 in/out degree = 0.11 for that sub-graph. After sorting the graph, the position of our original graph is the first, hence p-value $p1 =$ 0.5% (most significant pathway given the 200 graphs). In Sample 2, the mapped enzymes (filled rectangles) form a sub-graph with density = 1.375, number of 0 in/out degree = 0.22 for that sub-graph and its position after sorting is 148. This results in a p-value $p2 = 74.5\%$ (not a significant mapping).

Based on the value of p1 and p2, $probDiff(ko00062) = .74$ which is greater than 50%. We conclude that ko00062 is differentially significant in the two samples.

Results and Discussion

Data preparation

Bugula neritina is a colonial marine invertebrate found in temperate waters around the world [11]. *B. neritina* associates with an uncultured microbial symbiont, "*Candidatus* Endobugula sertula" [12] that has been shown to produce bryostatins, bioactive compounds that protect the host larvae from predation [13, 14]. Because of the pharmaceutical potential of the bryostatins, and the inablilty to grow the symbiont in the laboratory, we chose to examine host gene expression in the presence and absence of the symbiont to understand more about the molecular underpinnings of this relationship. In addition, as symbiont endow the larvae with high concentrations of bryostatins compared to the adult [15, 16], we also wanted to examine host gene expression in portions of the colony that possess reproductive structures termed ovicells, and those without ovicells.

Adult colonies of *B. neritina* growing on floating docks were collected from four locations on the Eastern coast of the USA: Radio Island Marina and Yacht Basin Marina, Morehead City (North Carolina) in March 2012, and Oyster public docks, Oyster (Virginia) in June 2012. The colonies were rinsed in filtered sea water and preserved in TRIzol reagent (Invitrogen, Carlsbad, CA) at -80 degree celsius prior to RNA extraction. Total RNA was then extracted from the preserved samples (RNeasy Mini kit, Qiagen,Inc., Valencia, CA, USA). The RNA was purified and treated with RNase-free DNaseI to remove any contaminating DNA. The purified total RNA was processed according to standard operating procedure for preparation of mRNA library for sequencing (TruSeq RNA Sample Preparation Kit, Illumina, San Diego, CA, USA). The adapter-ligated cDNA library was hybridized to the surface of Illumina flow cell and sequenced on an Illumina HiSeq 2500 sequencing platform. The paired-end reads were assembled de novo using Trinity software package [17] and the assembled contigs were annotated by performing blastx searches (Translated Query-Protein Subject BLAST 2.2.26+) against the Swiss-Prot database.

The samples used for analysis include: Lane 1: symbiotic; Lane 2: non symbiotic (aposymbiotic) [18]; Lane 3: symbiotic, with ovicells; and Lane 4: symbiotic, without ovicells. The symbiotic relationship was assessed in the collected colonies by PCR analysis for the presence of the bryostatin biosynthetic gene cluster gene, bryS [18, 19]. We report only the values averaged over all 4 lanes. The reads were assembled into contigs by Trinity. We also conducted a Trinity assembly on the union of all reads - about 221,818,850 of them - 2x50 bp using 22 Gb of space. We obtained 166,951 contigs, after filtering with RSEM-isopct-cutoff=1.00, also 76,769 ORFs, 37,026 BLAST hits of translated ORFs against the SwissProt database and around 12,748 proteins hits. This translates to 59.37% ORFs hits and 63.35% contigs hits. Using IsoDE, we were able to identify 1485 differential expressed genes between the two different conditions, the symbiotic and aposymbiotic *B. neritina*.

Pathway extraction and graph generation

By *de novo* co-assembly of RNA-Seq data and BLAST-ing resulting contigs against protein databases, with a certain confidence, we can infer the ortholog groups expressed in the sample. This is an important attribute of KEGG. We use KEGG to generate pathways from Trinity contigs and proteins. From the pathway databases we can easily extract the enzyme information associated with each pathway. We actually extracted all pathways along with all mapped proteins.

KEGG represents proteins as KO numbers and we also follow this representation. It uses KGML, an exchange format of KEGG pathway maps, to interact with external applications.The next step was to download all KGML files associated with the pathways using the API provided by KEGG. To convert KGML files to graph of node and vertices, we implemented and ran a novel tool called KGMLPathway2Graph. Mapping the output of KGMLPathway2Graph with KO numbers from KEGG analysis of our data, allowed us to compute pathway significance through p-$value$.

Results

Pathway expression differences in symbiotic and aposymbiotic *Bugula neritina* (Lanes 1 and 2) are shown in Table 1 and in Table 2 respectively following vertex labels permutation and the edges permutation models. We ran our graph-based models algorithm 1 with $m = 200$ graphs each generated after $n = 1000000$ permutations of labels or edges. This high number of permutations is necessary to introduce sufficient differences in the generated graphs. Table 3 presents the results on differential analysis of pathway activities between the same two *Bugula neritina* conditions.

In Additional file 1, we present a summary of transcripts differential expression (DE) analysis results using IsoDE [9] and pathway activity inference results.

Validation

From our statistical analysis, we identified some pathways that were differentially expressed (DE) by all methods. The next step was to experimentally validate these results. The first validation step is done through IsoDE, a software to analyze differentially expressed genes. Through KEGG, we are able to get all the proteins (contigs) participating in a pathway. IsoDE then indicates which of those contigs are also DE. From those DE contigs, we extracted the genes that will be tested via quantitative PCR (qPCR) experiment, the next validation step.

The goal of qPCR is to quantify the level of expression in the symbiotic and aposymbiotic *B. neritina.* It is used to validate the gene expression given by IsoDE. Primers were designed using Primer 3 Plus [20]. Total RNA from recently collected symbiotic and aposymbiotic *B. neritina* colonies was converted to cDNA using Superscript III (Invitrogen, Carlsbad, CA, USA) using random hexamers according to the manufacturer's instructions. cDNA was subjected to qPCR analysis after and expression in the samples was compared using the $2\triangle Ct$ method [21]. The reference gene used with the glyceraldehyde-3-phosphate dehydrogenase gene, a housekeeping gene identified from the *B. neritina* transcriptome (contig m.4423) [22, 23]. The expression of three genes identified as being differentially expressed in symbiotic and aposymbiotic animals from the fatty acid elongation pathway (ko00062) were compared.

Using IsoDE, nine gene pathways were chosen from KEGG and 2485 top differentially expressed contigs were taken from the list of all contigs. Within the selected pathways, there was a total of 637 contigs extracted. Each gene in these contigs was checked for fold change 1.2 or higher. Next, the number of genes that had significant fold change was compared to the total number of genes in the pathway. If the number of genes with fold change of 1.2 or higher account for at least 15% of the total number of genes in that pathway, then this pathway was chosen to be further investigated. The pathways chosen are fatty acid elongation (ko00062), peroxisome (ko04146), ribosome biogenesis in eukaryotes (ko03008), RNA transport (ko03013) and drug metabolism - other enzymes (ko00983) [3].

The fatty acid elongation pathway contains fourteen (14) KEGG mapped contigs and three (3) of those were found significantly differentially expressed. They are very-long-chain 3-oxoacyl-CoA reductase (DHB12), 3-ketoacyl-CoA thiolase (fadA) and 3-ketoacyl-CoA thiolase B, peroxisomal (THIKB). Once all the contigs in the pathway were checked, additional information was compiled: KEGG pathway and protein numbers, contig number, UniProt accession number and predicted fold change between symbiotic and aposymbiotic *B. neritina* [24]. qPCR primers were designed using Primer 3 Plus and ordered from Integrated DNA Technology [20]. The primers were tested using cDNA at concentrations from 1 $ng/\mu L$ to 0.1 $pg/\mu L$. In order to use $\triangle\triangle Ct$ method, every primer had to have the same efficiency and efficiency around 100% [21].

RNA was extracted from symbiotic and aposymbiotic *B. neritina* colonies. Following the Direct-zol RNA MiniPrep protocol (Zymo Research Corp., Irvine, California, USA), 50 mg of *B. neritina* tissue was homogenized and RNA was extracted. Then the RNA was further purified using the OneStep PCR Inhibitor Removal Kit (Zymo Research Corp., Irvine, California, USA). To eliminate any contaminating genomic DNA, a DNase I treatment was performed according to the manufacturer's protocol (DNase I Recombinant, RNase-free, Roche, Mannheim, Germany). Finally, the RNA was further purified with the RNA Clean-up and Concentrator Kit (Zymo Research Corp., Irvine, California, USA). The concentration of RNA was quantified in triplicate using a Nanodrop spectrophotometer.

Both symbiotic and aposymbiotic cDNA were synthesized using the Superscript III protocol (Superscript III First Strand Synthesis System for RT-PCR, Invitrogen by Life Technologies, Carlsbad, California, USA) with random hexamers. qPCR primer efficiencies were determined using qPCR. All qPCR reactions were performed using 7500 Fast Real-Time PCR system (Applied Biosystems) with hot-start Taq polymerase, SYBR Green fluorescent dye and ROX passive reference dye (Maxima SYBR Green/ROX qPCR Master Mix (2X), Life Technologies, Carlsbad, California, USA). The efficiencies for each primer pair were calculated using the slope of amplification curve in the equation E = 10(-1/slope) [25].

The expression levels of three genes, *fadA*, DHB12, and THIKB were measured in symbiotic and aposymbiotic *B. neritina* with G3P acting as an endogenous control. The reactions were run in triplicate for symbiotic and aposymbiotic cDNA along with a negative template control. The Ct averages and standard deviations were calculated to find the Ct differences between the target gene and the control ($\triangle$Ct) and $\triangle$Ct standard deviation. $\triangle\triangle$Ct was calculated by subtracting the symbiotic or aposymbiotic $\triangle$Ct by the symbiotic $\triangle$Ct. This resulted in symbiotic $\triangle\triangle$Ct equal to 0 to compared the fold change between symbiotic and aposymbiotic expression levels.

As presented in Table 4, the fold change predicted by differential expression analysis, using IsoDE, for these three genes indicated that expression was higher in the aposymbiotic B. neritina. *fadA* had a predicted fold change of 2.91, while DHB12 had a value of 1.90 (non-significant difference), and THIKB equaled 2.84. qPCR analysis showed that when aposymbiotic gene expression was compared to symbiotic gene expression, *fadA* had 6.88 higher expression in aposymbiotic *B. neritina*. DHB12 had 0.66 times lower expression and THIKB had 2.52 higher expression, indicating that computational method closely predicted expression in independent biological samples.

Discussion

Although all the EM and the graph-based methods worked on the same data generated by KEGG, the input to each approach were very different. For example, the Trinity output of sample1 on KEGG generates about 306 pathways. All of these pathways were considered for EM methods while only a small subset of 80 was used as input to each of the graph-based model. Different factors contributed to this reduced number of pathways been analyzed in the edge/vertex swapping model: (1) We were not able to extract the KGML of all pathways from from KEGG; (2) We were not able to convert all KGML to actual graph and (3) Some graphs did not carry enough mapping to be significant (we excluded pathways with less than 3 ortholog groups mapped).

Consequently, the graph-based approaches yield considerable less differentially expressed pathways than EM methods although results from both models in the graph-based approaches were very consistent. Also, the graph-based analysis appears to be more stringent selecting only the pathways which are the farthest apart according to our statistic criteria.

Looking at the overall small number of differential pathways, we can say that Bugula n. with or without the symbiotic relationship still exhibits very similar reaction. As shown in Table 5, the over all difference in the pathway activity and differentially expressed transcripts between the two samples is very small.

Because working with non-model organisms, such as *B. neritina*, is more challenging due to the lack of tools for genetic manipulation, for future works, we plan on the following. First, run XPathways tools in other organisms, including model organisms to further verify their efficacy and second, extend the model to handle not only metabolic pathways put also signaling pathways.

Conclusions

XPathway tools with its two computational approaches is able to efficiently infer pathway activity as well as pathway significance, respectively an expectation maximization and a graph-based approach. Rather than trying directly to identify differentially expressed genes from RNA-Seq data for a non-model organism, XPathway tools allows to more accurately predict differential expression of genes using wealth of information collected in the KEGG database for related organisms. Our experimental comparisons on *Bugula neritina* RNA-Seq data with or without the symbiotic bacteria enable the detection and comparison of pathways with metabolic difference. qPCR experiment successfully validated our findings.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

XPathway tools have been implemented in C and Python and can be run on any platform with Python or C compiler on the machine. KEGG database is publicly available. The source code and installation instructions are available at http://alan.cs.gsu.edu/NGS/?q=content/xpathway.

Competing interests

Authors declare that they have no competing interests.

Author's contributions

YTT prepared the testing data used to generate the experimental results, implemented and tested the algorithms involved in XPathway. YTT also contributed to designing the algorithms and wrote part of the manuscript. SS ran the IsoDE to compute transcript differential expression and wrote part of the manuscript. QC developed, implemented and tested the algorithms for converting KEGG pathway to graph. MM, KB and AR implemented and ran the qPCR experiment, from the selection of the genes to the lab analysis. Igor M. and OG participated in validating the XPathway algorithms. Ion M. ran Trinity and Blast and also contributed to designing the algorithms. NL provided the data and designed and supervised the qPCR analysis. NL also wrote part of the manuscript. AZ contributed to designing the algorithms and the experiments, writing the manuscript and supervised the project. All authors read and approved the final manuscript.

Acknowledgements

This work has been partially supported by the IIS-0916948 from NSF, a Collaborative Research Grant from Life Technologies, and the Molecular Basis of Disease Area of Focus Georgia State University.

References

1. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC: **Integrative analysis of environmental sequences using MEGAN4**. *Genome research* 2011, **21**(9):1552–1560.
2. Konwar KM, Hanson NW, Pagé AP, Hallam SJ: **MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information**. *BMC bioinformatics* 2013, **14**:202.
3. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic acids research* 2000, **28**:27–30.
4. Ye Y, Doak TG: **A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes**. *PLoS computational biology* 2009, **5**(8):e1000465.
5. Sharon I, Bercovici S, Pinter RY, Shlomi T: **Pathway-based functional analysis of metagenomes**. *Journal of Computational Biology* 2011, **18**(3):495–505.
6. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al.: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545–15550.
7. Efron B, Tibshirani R: **On testing the significance of sets of genes**. *The annals of applied statistics* 2007, :107–129.
8. Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, Voichita C, Drăghici S: **Methods and approaches in the topology-based analysis of biological pathways**. *Frontiers in physiology* 2013, **4**.
9. Al Seesi S, Tiagueu YT, Zelikovsky A, Măndoiu II: **Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates**. *BMC genomics* 2014, **15**(Suppl 8):S2.
10. Cheng Q, Zelikovsky A: **Combinatorial Optimization Algorithms for Metabolic Networks Alignments and Their Applications**. *IJKDB* 2011, **2**:1–23.
11. Trindade-Silva AE, Lim-Fong GE, Sharp KH, Haygood MG: **Bryostatins: biological context and biotechnological prospects**. *Current opinion in biotechnology* 2010, **21**(6):834–842.
12. Haygood MG, Davidson SK: **Small-subunit rRNA genes and in situ hybridization with oligonucleotides specific for the bacterial symbionts in the larvae of the bryozoan Bugula neritina and proposal of" Candidatus endobugula sertula"**. *Applied and environmental microbiology* 1997, **63**(11):4612–4616.
13. Davidson S, Allen S, Lim G, Anderson C, Haygood M: **Evidence for the Biosynthesis of Bryostatins by the Bacterial Symbiont "Candidatus Endobugula sertula" of the BryozoanBugula neritina**. *Applied and Environmental Microbiology* 2001, **67**(10):4531–4537.
14. Lopanik N, Lindquist N, Targett N: **Potent cytotoxins produced by a microbial symbiont protect host larvae from predation**. *Oecologia* 2004, **139**:131–139.
15. Lindquist N, Hay ME: **Palatability and chemical defense of marine invertebrate larvae**. *Ecological Monographs* 1996, :431–450.
16. Lopanik NB, Targett NM, Lindquist N: **Ontogeny of a symbiont-produced chemical defense in Bugula neritina (Bryozoa)**. *MARINE ECOLOGY-PROGRESS SERIES-* 2006, **327**:183.
17. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al.: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nature Biotechnology* 2011, **29**(7):644–652.
18. Linneman J, Paulus D, Lim-Fong G, Lopanik NB: **Latitudinal Variation of a Defensive Symbiosis in the Bugula neritina (Bryozoa) Sibling Species Complex** 2014.
19. Sudek S, Lopanik NB, Waggoner LE, Hildebrand M, Anderson C, Liu H, Patel A, Sherman DH, Haygood MG: **Identification of the putative bryostatin polyketide synthase gene cluster from "Candidatus Endobugula sertula", the uncultivated microbial symbiont of the marine bryozoan Bugula neritina**. *Journal of natural products* 2007, **70**:67–74.
20. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA: **Primer3Plus, an enhanced web interface to Primer3**. *Nucleic acids research* 2007, **35**(suppl 2):W71–W74.

21. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2- ΔΔCT method**. *methods* 2001, **25**(4):402–408.

22. Lee PD, Sladek R, Greenwood CM, Hudson TJ: **Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies**. *Genome Research* 2002, **12**(2):292–297.

23. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E: **Housekeeping genes as internal standards: use and limits**. *Journal of biotechnology* 1999, **75**(2):291–295.

24. Magrane M, Consortium U, et al.: **UniProt Knowledgebase: a hub of integrated protein data**. *Database* 2011, **2011**:bar009.

25. Pfaffl MW: **A new mathematical model for relative quantification in real-time RT–PCR**. *Nucleic acids research* 2001, **29**(9):e45–e45.

Figures

Figure 1: XPathway analysis flow. The branches represent the two approaches used to compute pathway significance in the case of graph-based on the left and pathway activity level in the case of the expectation maximization approach on the right. Both methods are validated by computing contigs/transcripts differential expressions and qPCR as the last step of the flow.

Figure 2: Expectation maximization approach to compute pathway activity. This bipartite graph consists of a set A representing reads/contigs/ORF/proteins and the set B is for ORFs/proteins/ortholog groups/EC (Enzyme Commission) numbers. The arcs represent mapping between elements of both sets. For our binary EM, the set A consists of contigs mapped to ortholog groups and the weight of each arc is 1.

Figure 3: Vertex labels swapping model for random graph generation. We only swap vertices which have different labels. A label is an attribute of a vertex representing a mapped or not protein.

Figure 4: Edge swapping model for random graph generation. Before swapping the edges, we check that the in and out-degree of the vertices involved remain the same.

Figure 5: Pathway differential analysis. In sample 1, the mapped enzymes (filled rectangles) form a sub-graph with density = 1.475, the number of 0 in/out degree = 0.11 and p-value=0.5. In Sample 2, the mapped enzymes (filled rectangles) form a sub-graph with density = 1.375, the number of 0 in/out degree = 0.22 and p-value=.74. Based on these p-value, we say that this pathways is differentially significant.

Tables

Table 1: Vertex label permutation: the p-values of pathways are computed from the symbiotic (Lane 1) and aposymbiotic (Lane 2) *B. neritina* data. This table presents the most significant divergence in pathway results, using the criteria described in the Methods section, they are declared differentially significant.

Pathway	p-value from symbiotic Bugula	p-value from aposymbiotic Bugula	ProbDiff
ko04146	.99	.05	.94
ko03008	.99	.05	.94
ko03013	.99	.05	.94
ko00983	.99	.05	.94
ko04530	.99	.05	.94
ko00062	.01	.75	.74
ko00400	.01	.99	.98
ko00071	.99	.01	.98
ko00100	.99	.01	.98
ko00910	.04	.99	.95
ko04122	.99	.03	.97
ko04713	.99	.01	.99

Table 2: Edge permutation: the p-values of pathways are computed from the symbiotic (Lane 1) and aposymbiotic (Lane 2) Bugula data. This table presents the most significant divergence in pathway results, using the criteria described in the Methods section, they are declared differentially significant.

Pathway	p-value from symbiotic Bugula	p-value from aposymbiotic Bugula	ProbDiff
ko04146	.99	.05	.94
ko03008	.99	.05	.94
ko03013	.99	.05	.94
ko00983	.99	.05	.94
ko04530	.99	.05	.94
ko00400	.01	.99	.98
ko04122	.99	.03	.97
ko04713	.99	.01	.99
ko00130	.01	.75	.74
ko00120	.01	.99	.98
ko00072	.99	.01	.98
ko00120	.99	.01	.98
ko00230	.04	.99	.95
ko00627	.99	.03	.97
ko00770	.99	.01	.99
ko00980	.99	.03	.97
ko04630	.99	.01	.99

Table 3: Pathway activities levels with ratio. Expression represents the expression level of the pathway activity in symbiotic (Lane 1) and aposymbiotic (Lane 2) *B. neritina* data. This table presents pathways with a ratio of 1.5 or higher in their activity level or pathways with a ratio of 0.66 or lower from the opposite direction. Using the criteria described in section 2, they are found to significantly differ in activities level.

Pathway	**Expression from- symbiotic Bugula**	**Expression from- aposymbiotic Bugula**	**Ratio between- pathway expressions**
ko00300	2.75	0.38	7.27
ko00290	4.26	1.71	2.49
ko04712	1.77	0.77	2.30
ko00903	1.88	0.84	2.25
ko01220	2.24	1.10	2.04
ko00981	2.00	1.00	2.00
ko04744	3.55	1.82	1.95
ko00626	1.48	0.76	1.93
ko00624	1.17	0.67	1.76
ko00072	2.17	1.25	1.74
ko00730	2.50	1.50	1.67
ko04730	3.80	2.40	1.58
ko00363	2.92	1.88	1.56
ko05150	2.13	1.38	1.55
ko04112	3.00	2.00	1.50
ko05219	1.67	2.50	0.66
ko00625	1.00	1.98	0.50
ko00984	1.00	2.00	0.50
ko00592	1.25	3.02	0.42
ko00965	0.66	1.66	0.40
ko00940	1.42	3.71	0.38
ko00460	0.60	2.02	0.30
ko00944	0.17	1.17	0.14

Table 4: Experimental quantification of fatty acid elongation gene expression by qPCR in symbiotic and naturally aposymbiotic *B. neritina*.

Genes	fadA	DHB12	THIK
Fold change of gene expression in aposymbiotic *B. neritina* compared to symbiotic by FPKM analysis	2.91	1.90	2.84
Gene expression in symbiotic *B. neritina* by qPCR analysis	2.46	4.34	4.34
Gene expression in aposymbiotic *B. neritina* by qPCR analysis	29.32	2.85	10.95
Fold change of gene expression in aposymbiotic *B. neritina* compared to symbiotic by qPCR analysis	6.88	0.66	2.52

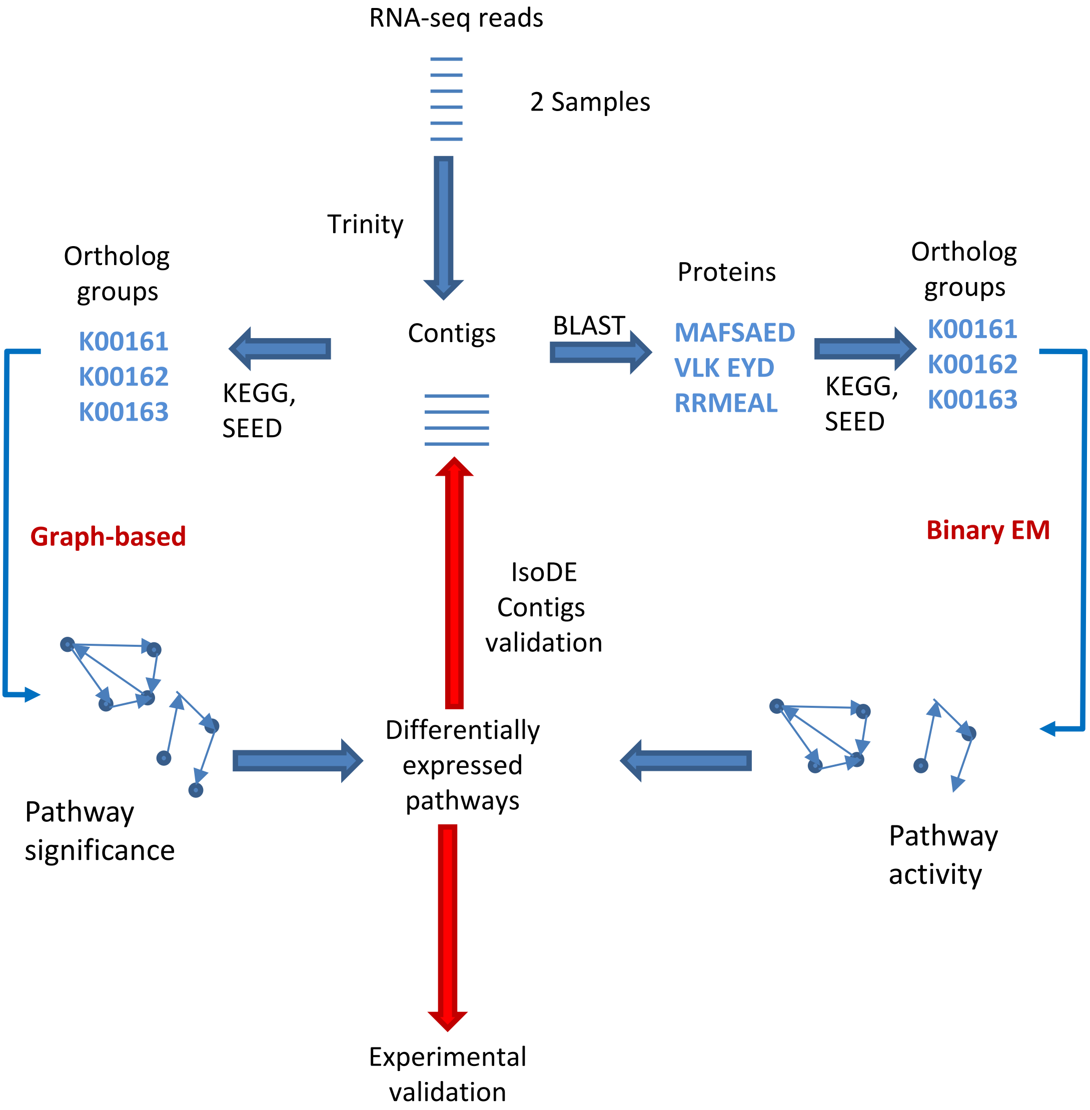
Table 5: Percentage of differentially expressed contigs with fold change (FC) of 2 and 1.5 respectively.

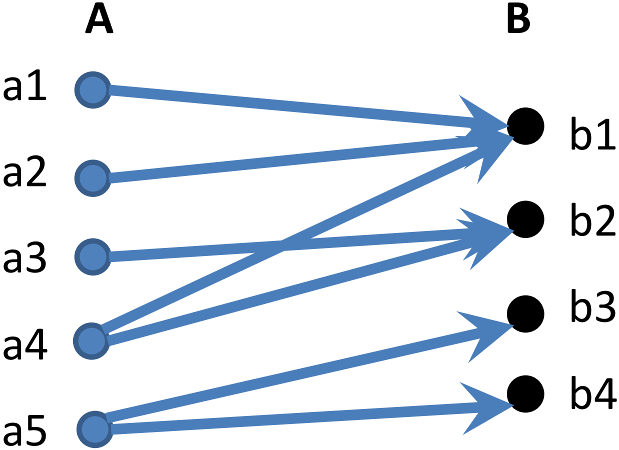
	FC = 2	**FC = 1.5**
Pathway	8%	12%
Contigs	13%	28%

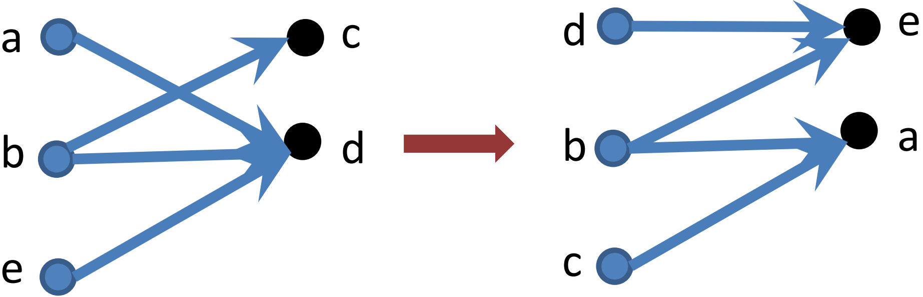
Additional Files

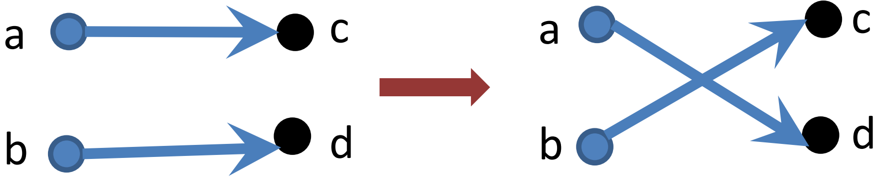
Additional file 1

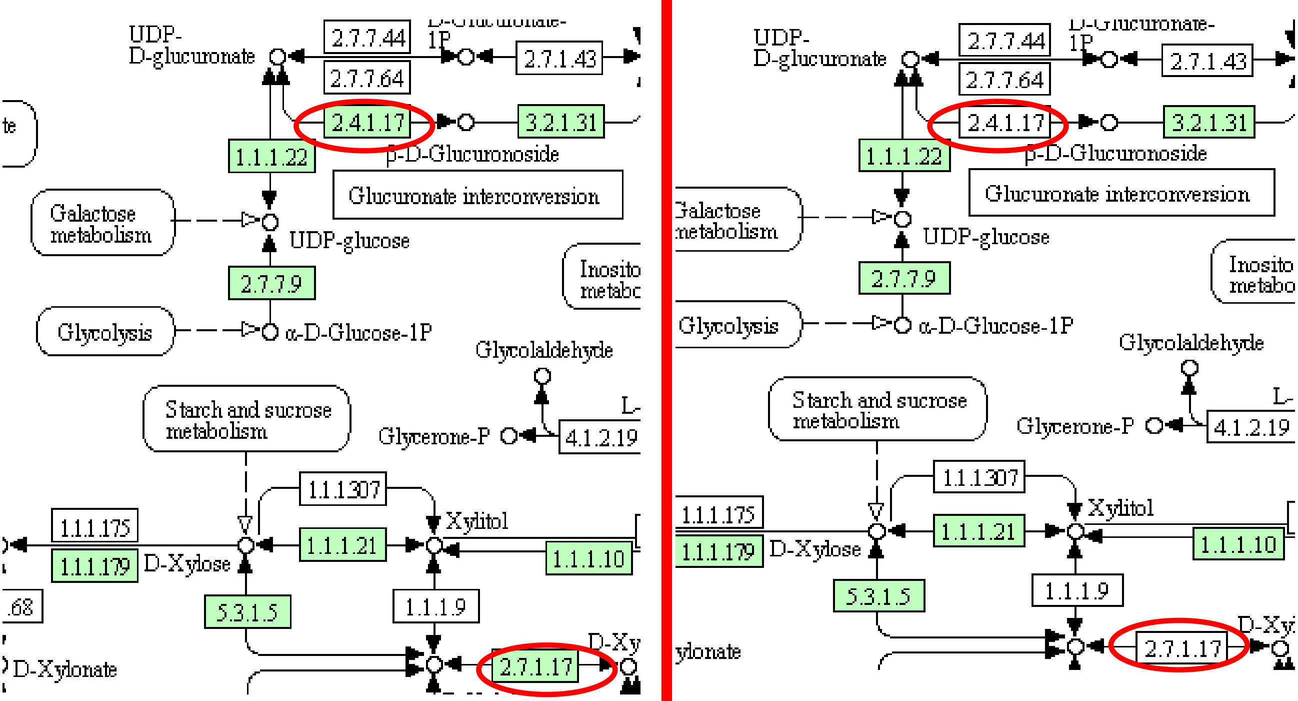
The Supplementary table is supplied in PDF format.


RNA-seq reads
2 Samples
Trinity
Ortholog groups
K00161
K00162
K00163
Contigs
BLAST
Proteins
MAFSAED
VLK EYD
RRMEAL
Ortholog groups
K00161
K00162
K00163
KEGG, SEED
KEGG, SEED
Graph-based
Binary EM
IsoDE
Contigs validation
Differentially expressed pathways
Pathway significance
Pathway activity
Experimental validation


A
B
a1
a2
a3
a4
a5
b1
b2
b3
b4


a
c
b
d
e
d
e
b
a
c


a
b
c
d
a
b
c
d



Sample 1

Sample 2

Supplementary table

Table S1 presents a summary of selection of genes for qPCR experiment. only pathway with a ratio of DE transcript of 15% or more were selected.

Pathway	Mapped-contigs	DE-contigs	Ratio-of-DE
ko00062:Fatty acid elongation	14	3	21.43%
ko00100:Steroid biosynthesis	8	1	12.50%
ko00250:Alanine, aspartate and glutamate metabolism	39	4	10.26%
ko04146:Peroxisome	98	15	15.31%
ko03008: Ribosome biogenesis in eukaryotes	67	10	14.93%
ko03013: RNA transport	148	22	14.86%
ko00983:Drug metabolism - other enzymes	28	4	14.29%
ko04530:Tight junction	237	15	6.33%

Table S1: DE contigs from selected pathways.