Available online at www.sciencedirect.com
ScienceDirect
J. Genet. Genomics 37 (2010) 205−218
JOURNAL OF GENETICS AND GENOMICS
www.jgenetgenomics.org

A gene family-based method for interspecies comparisons of sequencing-based transcriptomes and its use in environmental adaptation analysis

Zuozhou Chen [a], Hua Ye [a], Longhai Zhou [a], Chi-Hing C. Cheng [b], Liangbiao Chen [a, *]

[a] *Key Laboratory of Molecular and Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China*

[b] *Department of Animal Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

Received for publication 9 November 2009; revised 20 January 2010; accepted 3 February 2010

Abstract

We describe a new method for sequencing-based cross-species transcriptome comparisons and define a new metric for evaluating gene expression across species using protein-coding families as units of comparison. Using this measure transcriptomes from different species were evaluated by mapping them to gene families and integrating the mapping results with expression data. Statistical tests were applied to the transcriptome evaluation results to identify differentially expressed families. A Perl program named Pro-Diff was compiled to implement this method. To evaluate the method and provide an example of its use, two liver EST transcriptomes from two closely related fish that live in different temperature zones were compared. One EST library was from a recent sequencing project of *Dissosticus mawsoni*, a fish that lives in cold Antarctic sea waters, while the other was newly sequenced data (available at: http://www.fishgenome.org/polarbank/) from *Notothenia angustata*, a species that lives in temperate near-shore water of southern New Zealand. Results from the comparison were consistent with results inferred from phenotype differences and also with our previously published Gene Ontology-based method. The Pro-Diff program and operation manual can be downloaded from: http://www.fishgenome.org/download/Prodiff.rar.

Keywords: transcriptome comparison; EST; protein family; reference gene set

Introduction

Cross-species transcriptome comparison

In this post-genomic era, genome-wide transcription analyses are widely adopted, and comparative transcriptomics greatly accelerated the understanding of the relationship between transcriptome differentiation and phenotypic evolution. Researches currently carried out included transcriptome comparisons between human and non-human primates to explain the intellectual and behavioral differences between human and other primates (Enard et al., 2002; Caceres et al., 2003; Khaitovich et al., 2005b, 2006), comparison of transcriptomes from different populations adapted to different environmental conditions (e.g., temperature) (Whitehead and Crawford, 2006) and investigation of theoretical questions such as the neutrality of transcriptome evolution (Khaitovich et al., 2004, 2005a; Whitehead and Crawford, 2006; Hoffmann et al., 2007) or the relationship between transcriptome evolution and the evolution of genomic sequences (Lemos et al., 2005;

* Corresponding author. Tel & Fax: +86-10-6255 4807.
E-mail address: lbchen@genetics.ac.cn

DOI: 10.1016/S1673-8527(09)60039-4

Khaitovich et al., 2006; Gu and Su, 2007; Tirosh and Barkai, 2008).

How transcriptomes of an organism adapt to environment (e.g., low temperature) is one of the important fields of evolutionary transcriptomics. Transcriptome data from Antarctic notothenioid fish living in sub-zero sea waters and their related species from temperate waters provide excellent resources for the study of transcriptional adaptation mechanisms to geographical-time-scale temperature fluctuations. Recently，we generated more than 30,000 Expressed Sequence Tags (ESTs) from brain, liver, ovary, and head kidney tissues of *Dissostichus mawsoni* (*D. mawsoni*), a high Antarctic notothenioid species. Cross-species transcriptome comparisons between *D. mawsoni* and warm-water teleost fishes were conducted using a new method we designed to infer cold adaptation mechanisms (Chen et al., 2008). Here we describe the rational and evaluation of this method in detail and the utility of the computer program we implemented.

Sequencing-based approaches for cross-species transcriptome comparisons

The ability of microarray technology in monitoring the expression of tens of thousands of genes at the same time is exploited in almost all current transcriptome evolution studies to evaluate expression levels of a particular 'gene' across multiple species. Sequencing-based transcriptomes, however, are rarely used in cross-species transcriptome comparisons. Hybridization-based methods have limited application to cross-species transcriptome comparisons (Gilad et al., 2005, 2006) because of discrepancies arising from heterologous probes, probe-specific hybridization kinetics, and the restriction to a small number of model organisms because of lack of availability of commercial chips. Sequencing-based approaches, such as traditional Expressed Sequence Tag (EST), Serial Analysis of Gene Expression (SAGE) (van Ruissen and Baas, 2007)，Massively Parallel Signature Sequencing (MPSS) (Margulies et al., 2005) and Massively Parallel Pyrosequencing (Brenner et al., 2000), however, have many advantages.

In addition, sequencing-based methods have two more advantages for cross-species transcriptome comparisons. Firstly, as cross-species transcriptome comparisons usually require multiple transcriptomes from non-model organisms that do not have existing microarray platforms, sequencing-based methods provide a simple solution and are usually the first step to gain a transcriptional and genomic overview. Secondly, they share the same statistical Poisson distribution and data format, and sequencing-based transcriptomes from different sources are relatively easy to integrate and compare. For example, EST library data can be clustered and integrated into large databases in a single format, such as the NCBI UniGene and TIGR Gene Indices, greatly facilitating cross-comparisons of various EST libraries. Although the traditional EST sequencing approach is not as high-throughput as microarray technology, recently developed massively parallel sequencing technologies achieved high order magnitudes of throughput, which can be applied to obtain more robust, comparable and rich expression profile data, especially in cross-species comparisons (Meyers et al., 2004; Nobuta et al., 2007; Marioni et al., 2008; t Hoen et al., 2008).

With the development of massive parallel sequencing technology, and its high potential in evolution studies, methods for optimizing biological minings are highly demanded. Previously, gene expression patterns of sequencing-based transcriptomes are accurately modeled by the Poisson distribution (Audic and Claverie, 1997), and methods for analyzing such data employing the Chi-square test, Fisher's Exact test or the Audic-Claverie test (Audic and Claverie, 1997; Man et al., 2000) are well developed.

At present, however, we lack a systematic approach for the sequencing-based cross-species transcriptome comparison. Unlike transcriptome comparison within the same species, in which a set of common genes or transcripts can be used as references, and the expression level of each reference sequence can be uniformly evaluated between experimental samples, transcriptomes from different species do not share the same set of reference genomic sequences. Microarray-based cross-species transcriptome comparisons usually explicitly or implicitly employed a set of orthologous genes to form a reference set (Enard et al., 2002; Caceres et al., 2003; Liao and Zhang, 2006a, 2006b; Chen et al., 2007). There have only two attempts to overcome this problem for sequencing-based transcriptome comparison. The first attempt used reference gene pairs generated from the top BLAST (Altschul et al., 1990) hits of assembled ESTs of tomato and Arabidopsis (Fei et al., 2004). Such a 'top hit' approach is no better than ortholog approaches and suffers from the limitations as discussed below. In the second attempt, species-independent Gene Ontology (GO) (Ashburner et al., 2000) terms were used

as a reference set to be evaluated across species (Chen et al., 2006). However, such approach generates only the over- or under-represented GO terms without knowing which gene or gene family is involved, which is the information of the greatest importance for biologists.

A suitable and systematic approach for cross-species comparison of sequencing-based transcriptomes needs to solve at least two issues. The first is the nature of the orthologous gene relationship across species. As discussed above, a set of well-established orthologous genes suitable for evaluation in a one-to-one manner cross-species is required for cross-species comparisons, but very difficult or even impossible to obtain for a few obstacles. For example, when the divergence time between the two species in question increases, events of gene duplication, gene loss and divergence would lead to complex relationships between genes and become difficult to measure (Fig. 1A). In addition, EST sequences usually do not cover the full length of a gene, obtained sequences from different species might be too short to be identified as orthologs (Fig. 1B). Furthermore, in organisms without sequenced genome, the copy number of a gene is unknown and impossible to map its ESTs to the genome (Fig. 1C).

The second issue is that, under current EST clustering/assembly tools, Phrap, CAP3 (Huang and Madan, 1999) and TIGR assembler (Sutton et al., 1995), it is difficult to distinguish sequencing errors from closely related gene families (Liang et al., 2000) and whether two sequences are assembled or not is determined by program parameters. If one transcript is falsely split into two contigs or two transcripts are falsely assembled into one contig, the evaluation of gene expression of that transcript is inaccurate and leads to unreliable transcriptome comparisons (Fig. 1D).

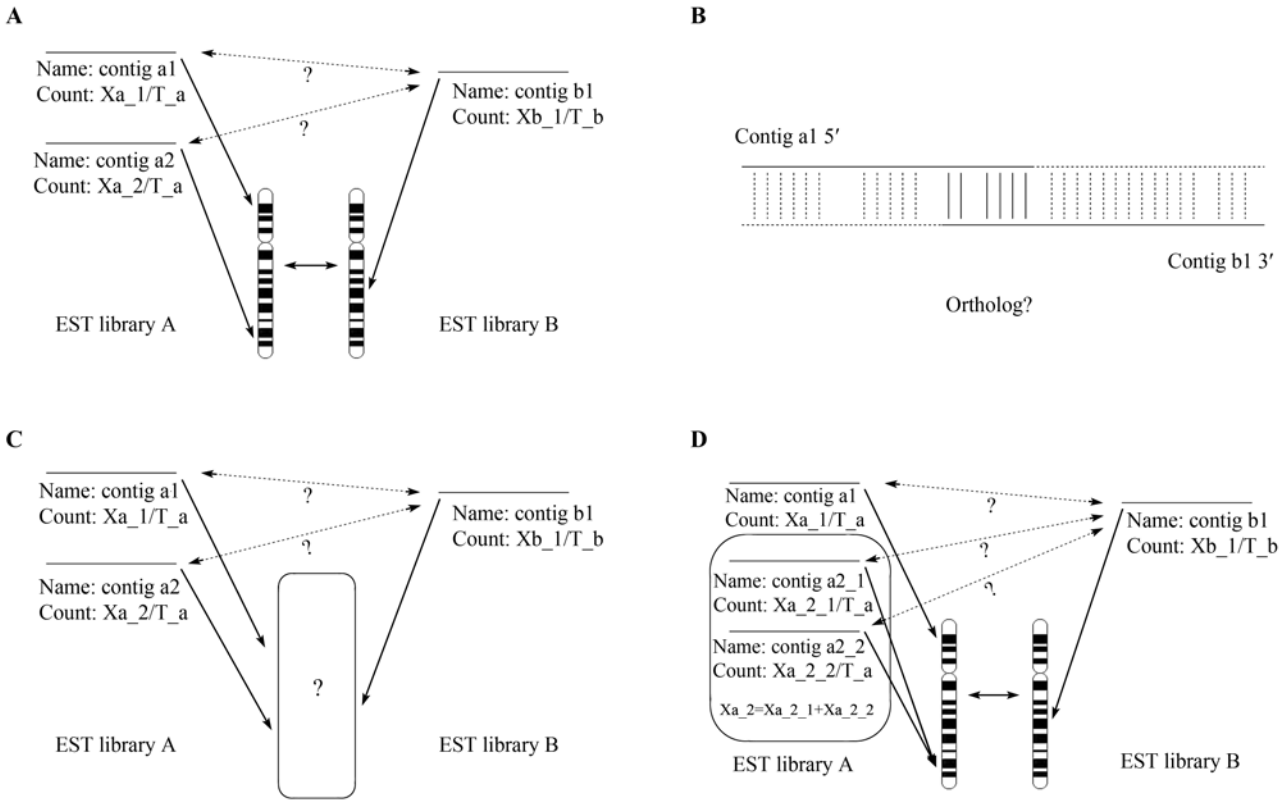

Fig. 1. Four possible problems associated with cross-species comparison of sequencing-based transcriptomes. **A:** the copy number of a gene may be different between two genomes, thus the object of comparison is ambiguous. **B:** the orthologous relationship is difficult to establish when the overlapping sequences are too short because of partial sequencing. **C:** comparison is difficult when the background genome is unclear. **D:** contig a2 is mistakenly spilt into two contigs by the assembly program. EST libraries A and B represent EST libraries of two species; contigs a1 and a2: example contigs in library A; contigs b1 and b2: example contigs in library B; Xa_1, Xa_2, and Xb_1: tag counts of contig a1, a2 and b1, respectively; T_a and T_b: the total tag counts of EST libraries A and B.

In light of the above two issues, here we propose a theoretical framework and a detailed method that uses protein-coding gene families as units for evaluating gene expression, instead of the traditional metric using a single pair of transcripts/genes. Using the metric proposed here, the expression level of gene families over multiple species can be calculated and compared by combining gene copy number and expression information. Our method can compensate for assembling errors arising from current clustering/assembling tools and also avoids the problem of ortholog identification. We have developed a program named Pro-Diff to implement this method.

We tested our method in two ways: first by comparing the liver transcriptomes of *D. mawsoni* and *N. angustata* to see if results obtained with this method agreed with biological expectations; second, results obtained were then compared quantitatively with the results previously generated using the GO-based method.

Materials and methods

The theoretical framework for cross-species comparisons among sequencing-based transcriptomes

Transcriptome comparison, intra- or inter-species, requires a set of stable objects (genes, gene families or GO terms) to be evaluated. Traditionally, in intra-species comparisons, researchers have used all the genes of a certain genome as a 'stable' objective set. However, the gene set of any given genome is in a constant state of flux. Even within one species, copy number variation is widely observed (Redon et al., 2006). Therefore, the stability of any object to be evaluated is not absolute but relative. The granularity of the compared-object should be increased to accommodate object variation, so that cross-species transcriptome comparisons can be implemented.

We use the word 'family' to encapsulate such an object. It contains two levels of meaning—broad and narrow ones. In the narrow sense, it is a family of protein coding genes defined by evolution or sequence similarity. In a broad sense, it can be a group of genes defined by any appropriate criterion, such as protein domains or functions. Although in this study, we concentrate on the narrowed definition of a 'family', our theoretical framework can apply for the broad definitions of 'family'.

Sequencing-based approaches generate partial sequences of a transcript named 'tags'—the EST. The tag count of a certain gene in a certain library represents the expression level of a gene. Here we use ESTs to represent all such tags generated from a single-pass sequencing effort. A flowchart for the method is shown in Fig. 2.

The family-wise expression level metric

We define the 'Tag Coverage Level of a protein-coding gene Family' (TCLF) as the total tag number of the contigs/clusters related to a particular gene family. For a certain gene family:

$$TCLF = f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i$$

Here xi is the tag count of a certain contig/cluster belonging to this family, and the TCLF is the sum of the expression level of all the members of the family.

This definition of TCLF brings in several advantages. By using a gene family as a unit to evaluate the gene expression level, the problems associated with ortholog-mapping and choosing the objects for evaluation are avoided. It also compensates for EST assembling errors, as whether two sequences are properly assembled or not, the final TCLF would be the same.

As larger libraries have larger TCLFs, the TCLF is normalized by the 'Total Tag Number assigned to protein-coding Families of each library' (TTNF) to give the 'Relative Tag Coverage Level of a protein-coding gene Family' (RTCLF):

$$RTCLF = TCLF/TTNF$$

The combination of RTCLFs of multiple gene families across multiple libraries is called a 'family profile'.

Here we use TTNF rather than 'total tag number of each library' to compensate for the heterogeneity of different libraries. Tags from different libraries may be of different quality lengths, or sequencing directions (3′ and 5′), and so the proportion of tags that can be assigned to protein-coding gene families will vary.

Family expression ratios and statistical testing

In order to reflect the differences in expression of each family between two transcriptomes, the 'Tag Coverage Ratio of a gene Family' (TCRF) can be calculated:

$$TCRF = RTCLF2/RTCLF1$$

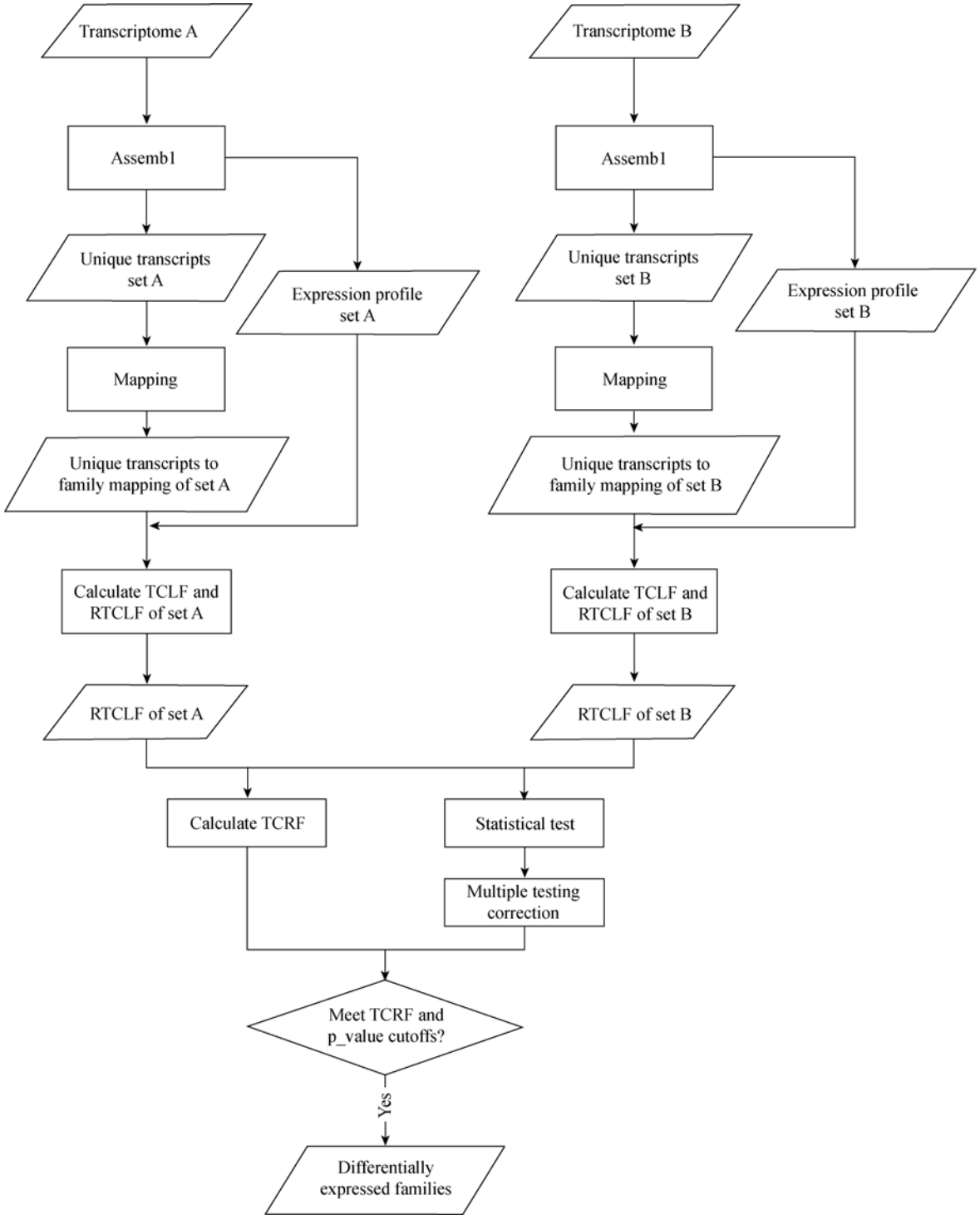

Fig. 2. Flowchart for family-wise comparison of two sequencing-based transcriptomes. Transcriptome is firstly assembled to obtain a unique transcript set and the expression profile of this set was generated. The unique transcripts are then mapped to a user-defined gene family set. The 'Tag Coverage Level of a protein-coding gene Family' (TCLF) and the 'Relative Tag Coverage Level of a protein-coding gene Family' (RTCLF) are calculated by integrating the mapping result and the expression profile. Using the TCLF and RTCLF of the two transcriptomes, the 'Tag Coverage Ratio of a gene Family' (TCRF) is calculated and statistical tests are implemented followed with a multiple testing correction. Finally, the families that meet the pre-defined criteria of a TCRF and a corrected p-value are determined as differentially expressed between the two transcriptomes.

For a comparison between two transcriptomes, the TCRF of each protein-coding gene family is calculated and each TCRF is then tested for significance. The statistical tests used are similar to tests implemented in intra-species comparisons in which the Chi-square test, Fisher's Exact test and Audic-Claverie test (Audic and Claverie, 1997) are widely applied (Man et al., 2000).

Here we conduct Fisher's Exact test for gene expression level of each family in the two transcriptomes of different species. First, a two by two contingency table is created (Table 1). Here a/(a+c) and b/(b+d) are RTCLFs of the two transcriptomes of a certain family. The marginal sums of (a+b), (c+d), (a+c) and (b+d) are fixed, and a,b,c and d are variables. The probability of each combination of a, b, c and d is calculated with the formula below:

$$p = (a+b)!(c+d)!(a+c)!(b+d)!/a!b!c!d!(a+b+b+d)!$$

To reject the null hypothesis that RTCLF1 and RTCLF2 are equal, all the p-values of different combinations of a, b, c and d should be summed if they are smaller or equal to the observed p-value p_o.

$$p = \sum p \ (p <= p_o)$$

Here p_o is the observed p-value of our observed tag number a, b, c and d.

Multiple testing corrections are made based on the raw p-values of Fisher's Exact tests. With a list of sorted p-value generated from above tests for each family, the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) of False Discovery Rate (FDR) control is used for the correction. Detailed description of this procedure can be found in (Benjamini and Hochberg, 1995) and also online (http://en.wikipedia.org/wiki/False_discovery_rate), which uses a linear step-up procedure to control the expected proportion of incorrectly rejected null hypotheses (type I errors). Differentially expressed protein-coding gene families are then selected based on the TCRF and FDR cutoff values.

Mapping contigs/unigenes to protein-coding gene families

A set of gene families must be defined and the unique transcripts are mapped to the families to calculate TCLF. Researchers can choose their own gene family schemes and mapping can be done manually, automatically or by both, depending on personal requirements. Here we recommend a simple pipeline to define a set of gene families and map the transcripts onto them (Fig. 3). The following procedure is recommended but not absolutely required in our method and researchers can choose alternative procedures for mapping. For this reason, the procedure is not included in the implemented program.

Table 1
The two by two contingency table for Fisher's Exact test

	Transcriptome A	Transcriptome B	Sum
Family A	a	b	a+b
Not Family A	c	d	c+d
Sum	a+c	b+d	a+b+c+d

The procedure is composed of two parts: generating a reference protein set, in which each protein represents a family, and mapping the unique transcripts to this protein set. The unique transcripts of the two sets are first pooled and then BLAST searched against a pre-clustered protein database. We use a protein cluster as an approximation to a protein family, for it is relatively simple to define. Furthermore, with a clustered large database such as Uniref50, it is relatively easy to find representative proteins having enough sequence similarity with the transcripts waiting to be analyzed. Top BLASTX hits are selected as reference proteins if they meet the user-defined criteria of similarity and High-scoring Segment Pair (HSP) length. Then TBLASTX is run separately with the selected reference proteins against the unique transcript set A and B. All of the TBLASTX hits for a certain reference protein are assigned to that family if they meet the user-defined criteria. If one transcript is assigned to more than one family, the redundant alignments are removed leaving only the one with the highest similarity.

cDNA library construction and EST sequencing of N. angustata

A cDNA library of *N. angustata* liver tissue was constructed using the pCMV-Script XR cDNA Library Construction kit (Stratagene, USA). Plasmids were isolated using the AxyPrep Easy-96 Plasmid DNA Isolation kit (Axygen Biosciences, CA, USA). DNA sequencing from the 5′ end of each cDNA clone was performed using BigDye version 3.1 (Applied Biosystems Inc., USA), and resolved on an ABI 3730 machine.

A pipeline of several programs and scripts were applied with parameters determined by experience and optimization. PHRED (Ewing and Green, 1998) software was used in base-calling with a stringent PHRED score cutoff of 30

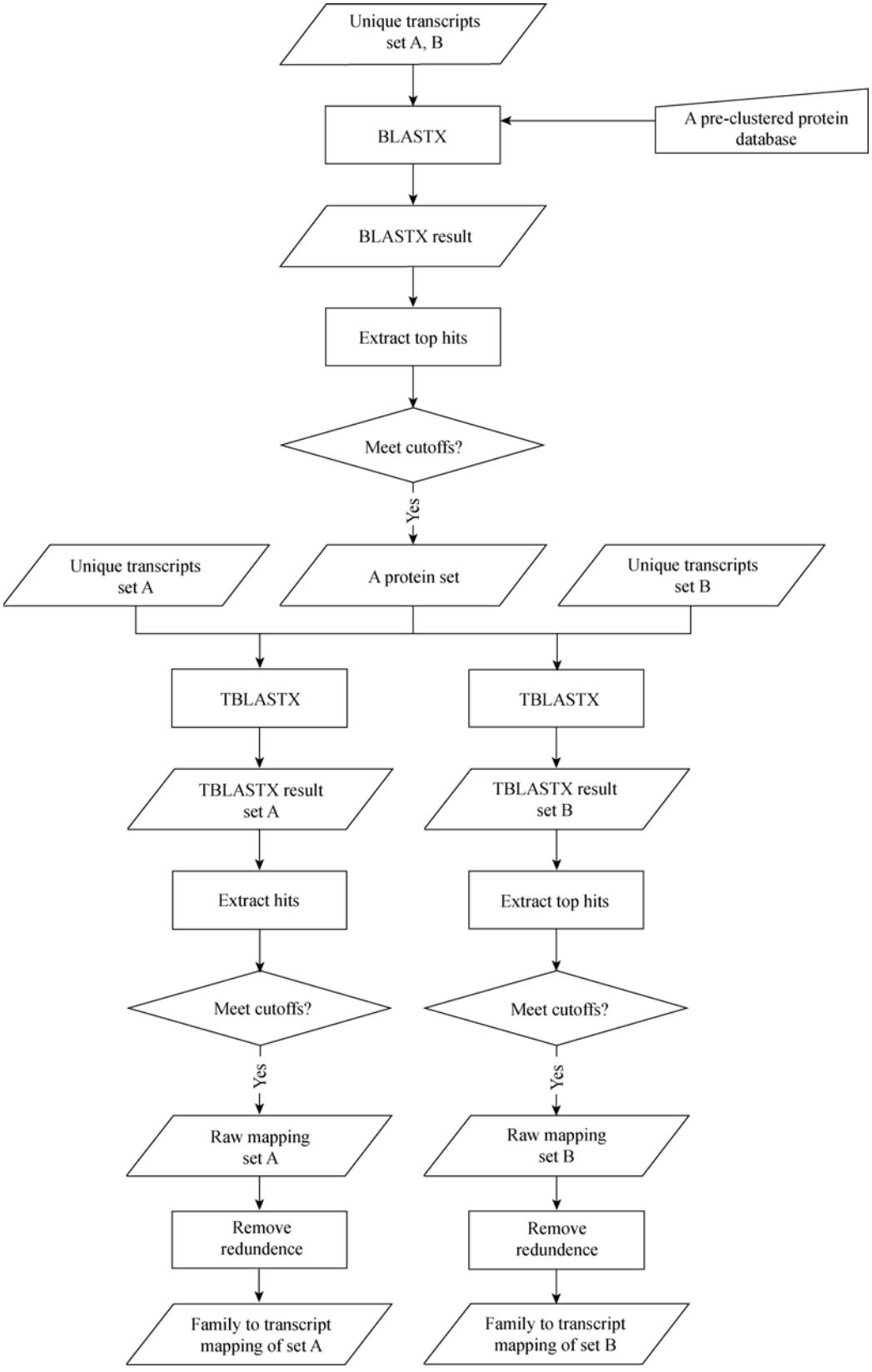

Fig. 3. The recommended procedure for mapping transcripts to protein-coding gene families.

for subsequent analyses, meaning 99.9% accuracy cutoff for each base pair. Vector and linker sequences were masked, and mitochondrial RNA, ribosomal RNA and contaminating *E. coli* genes were removed using an in-house pipeline compiled with the CROSS_MATCH, BLAST programs and Perl scripts. The cleaned sequences with lengths above 10 bp were submitted to dbEST (Boguski et al., 1993) and those exceeding 50 bp were used in the following analysis. ESTs were assembled into contigs using CAP3 software using the criteria of 50 bp overlap and similarity of 98%. The expression profiles of the contigs were extracted from the CAP3 assembly results with a Perl script.

Map EST assembly results for protein-coding gene families in D. mawsoni and N. angustata liver transcriptome comparisons

The files containing the *D. mawsoni* and *N. angustata* unique transcripts were searched using BLASTX against Uniref50 (Suzek et al., 2007) —a Uniprot (Wu et al., 2006) protein database pre-clustered by cd-hit (Whitehead and Crawford, 2006), using 50% similarity cutoff. The top BLASTX hit was selected if the alignment had greater than 60% identity and was more than 50 aa in length. Each Uniref50 protein was viewed as a representative of a protein-coding gene family, and was used as a reference set to evaluate family-wise expression levels across the liver transcriptomes of the two species. This reference protein set was then searched against the *D. mawsoni* and *N. angustata* unique transcript files using TBLASTX, and a transcript was mapped to a certain reference protein if the alignment had greater than 60% identity and was more than 50 aa in length. If a transcript mapped to multiple reference proteins, only the one with the highest alignment score was preserved.

GO over-representation analysis of the Pro-Diff results

Proteins that were representative of each gene family were searched against the Uniprot_sprot database (Wu et al., 2006) using BLASTP. GoPipe version 2 (Chen et al., 2005) was used to extract GO annotation from BLASTP results by searching the top BLASTP hit of each unique transcript against the GO annotation database of Uniprot (GOA). The GO annotations of the differentially expressed representative proteins of a gene family were extracted using a Perl script. The gene-count associated with each GO term (GO profile) for all the representative proteins and for the differentially expressed representative proteins were calculated by integrating the two GO annotations and the GO structure. The two GO profiles were then compared for each GO term using the Chi-square test and Fisher's Exact test followed by multiple testing correction using the linear step-up procedure. GO terms that met the criteria of having a ratio of $\geqslant 2$ or $\leqslant 1/2$ and FDR < 0.1 were chosen as over- or under-represented GO terms.

Results

A program for gene-family based interspecific transcriptome comparison

To implement the above method for cross-species transcriptome comparison, we developed a program named Pro-Diff, a freely available Perl program. This program is integrated into the GO-Diff package (Chen et al., 2006), which was formerly developed to screen over- or under-expressed genes grouped by GO terms for intra- or inter-species comparisons. Several parameters must be set before running the program, including TCRF cutoff, FDR cutoff, contig-family mapping file name, gene expression file name, library names and output file name. Detailed use and installation information can be found in the program manual on our website (http://www.fishgenome.org/bioinfo).

Cross-species comparison between transcriptomes from livers of D. mawsoni and N. angustata

To evaluate the use and validity of this method, we implemented this method in a comparison between liver transcriptomes of the Antarctic notothenioid, *D. mawsoni* that inhabits the coldest water of the world, and its close relative *N. angustata* that is in the same family of Nototheniidae but lives in cool-temperate waters of southern New Zealand.

A total of 5,492 ESTs were sequenced using adult *N. angustata* liver tissue, 4,068 of which passed a pipeline with stringent criteria. These sequences were further assembled into 1,056 unique transcript genes using the CAP3 program, and the expression profile was extracted from the assembly resultant file (sequences and annotation available at http://www.fishgenome.org/polarbank). The sequences and expression information of *D. mawsoni* liver was generated using the same pipeline and criteria as in

our other recently published research (Chen et al., 2008). The unique transcripts of *N. angustata* and *D. mawsoni* were then mapped to gene families represented by Uniref50 proteins. Pro-Diff was run to integrate these two mapping results and expression profiles and compare family-wise expression levels of the two transcriptomes. Results are shown in Table 2.

Results of the comparison are largely consistent with current knowledge of fish physiology, and show that the adaptive phenotypes of *D. mawsoni* can be explained with molecular data. For example, the skeleton of adult *D. mawsoni*, one of the few notothenioid species that are pelagic, has a low mineral content and contains considerable cartilage to reduce its body density allowing neutral buoyancy (Eastman and DeVries, 1981), in contrast to the hard-boned, shallow benthic *N. angustata*. The *D. mawsoni* phenotype described 27 years ago can now be explained by its high expression of fetuin genes (Table 2) which can inhibit precipitation of calcium phosphate (Heiss et al., 2003), and suppresses osteogenesis (Binkert et al., 1999). *D. mawsoni* also has large lipid deposits that provide static lift contributing to neutral buoyancy (Eastman and DeVries, 1982; Clarke et al., 1984). Previous analysis of plasma from *D. mawsoni* showed high levels of high density lipoprotein (Metcalf et al., 1999), which agrees well with our current finding of high expression of apolipoprotein in the liver (Table 2) that would facilitate lipid binding and transport. High expression of superoxide dismutase in *D. mawsoni* (Table 2) is consistent with potentially greater levels of ROS caused by increased O_2 solubility at low temperatures (Abele and Puntarulo, 2004) relative to temperate water conditions.

Interestingly, the zona pellucida sperm-binding protein 3 (ZP3/ZPC) gene family has an elevated expression in *D. mawsoni* liver, and EST assembly results show that there are at least 13 closely-related members in this family (Chen et al., 2008). This phenomenon is supported by a comparative genomic hybridization analysis and Southern blot of the *D. mawsoni* genome, which revealed that there are many more ZPC5 copies in *D. mawsoni* and other Antarctic notothenioids than in their non-Antarctic cousins. Multiple copies of ZP3 genes and their elevated expression levels provide a good example demonstrating that gene copy number determine the expression abundance of a gene family. Multiple sequence alignment (Fig. 4) of the peptides of the 6 contigs selected from the 13 unique transcript genes shows that peptides coded by different copies of ZP3 genes are highly conserved. In such a situation, evaluating total family-wise gene expression is more appropriate than considering gene expression on an individual basis.

Meta-analysis of the results of the comparisons

The comparison results can be affected by various factors. These factors can be environmental factors such as temperature, oxygen solubility or diet, or other factors such as species divergence, individual variation and sample sizes (tag number). To mine the environmental adaptation mechanisms of *D. mawsoni*, meta-analysis by comparing liver transcriptomes of *D. mawsoni* with liver tissues from many other species is required. Using meta-analysis, effects from other factors can be screened out and common themes in environmental adaptation are revealed. In our previous study, a liver transcriptome of *D. mawsoni* was compared to liver EST libraries from three other unrelated species, namely *Danio rerio*, *Fundulus heteroclitus*, and *Oryzias latipes*. Meta-analysis of these three previous comparisons showed that up-regulated genes were consistent, giving rise to an intersection of commonly up-regulated genes. Down-regulated genes, however, were not consistent across comparisons. This explains why 94% of the commonly differentially expressed gene families were up-regulated (Chen et al., 2008).

This phenomenon was confirmed further by adding our comparison of transcriptomes from the liver tissue of *D. mawsoni* and *N. angustata* to the previous meta-analysis described above (the last column of Table 2). Eleven out of 16 up-regulated gene families of *D. mawsoni* showed agreement with our previous study, and there were no contradictory results. Of the down-regulated gene families, however, the results for only one family were in agreement with the previous study, while 6 out of 35 families showed contradictory results (Table 2).

Cross-validation by comparing to the GO-Diff results

To the best of our knowledge there is so far only one tool (GO-Diff) other than Pro-Diff for conducting cross-species transcriptome comparisons with sequencing-based data. This tool was developed by us and is mainly used for mining functional differentiation between EST-based transcriptomes. As GO terms are species-independent in nature, it can also be applied in cross-species transcriptome comparisons.

Table 2
The differentially expressed protein-coding gene families represented by top hit Uniref50 proteins

Uniref50 protein ID	RTCLF_lib1	RTCLF_lib2	TCRF (lib2/lib1)	Corrected p-value	Description	Agreement
Q4SMM6	0	0.012	100	1.31E-06	Fetuin-B	–
Q4REC1	0	0.013	100	8.81E-08	Importin subunit alpha-1	–
Q4T7M2	0	0.020	100	4.32E-12	Zona pellucida sperm-binding protein 2	T
Q4SNR9	0	0.054	100	2.27E-36	Zona pellucida sperm-binding protein 3	T
Q9UBR2	0	0.015	100	6.64E-09	Cathepsin Z	T
Q4QY86	0	0.023	100	7.06E-15	Unknown	–
P27449	0	0.0054	100	0.017	Vacuolar ATP synthase 16 kDa proteolipid subunit	T
Q3ZE27	0	0.0062	100	0.0068	Zona pellucida sperm-binding protein 2	T
Q8AYL3	0	0.0078	100	0.00053	Zona pellucida sperm-binding protein 2	T
P08228	0	0.0051	100	0.028	Superoxide dismutase [Cu-Zn]	T
O42364	0	0.014	100	3.27E-08	Apolipoprotein Eb	T
P07900	0	0.0081	100	0.00033	HSP 90-alpha	–
P02679	0	0.0094	100	4.00E-05	Fibrinogen gamma chain	T
Q4SY35	0.00098	0.013	14	1.33E-05	Haptoglobin	T
Q2LK88	0.00098	0.0073	7.4	0.044	Fucolectin-4	T
P09972	0.014	0.0035	0.26	0.0023	Fructose-bisphosphate aldolase C	T
P08865	0.013	0.0027	0.21	0.0011	40S ribosomal protein SA	–
Q02878	0.0093	0.0019	0.20	0.010	60S ribosomal protein L6	F
P62701	0.0073	0.0013	0.18	0.038	40S ribosomal protein S4	–
Q3V5Y0	0.040	0.0073	0.18	7.06E-15	Lactose-binding lectin l-2	F
UPI00005A4635	0.018	0.0032	0.18	2.25E-06	Elongation factor 1-alpha 2	–
Q75UL8	0.047	0.0083	0.18	8.03E-18	Hemopexin	–
P80429	0.067	0.00941	0.14	2.78E-30	Serotransferrin-2	–
P62753	0.0073	0.00081	0.11	0.0041	40S ribosomal protein S6	–
Q4RUP8	0.0078	0.00081	0.10	0.0019	unknown	–
P98093	0.0059	0.00054	0.092	0.013	Complement C3-1	F
P49946	0.0064	0.00054	0.085	0.0061	Ferritin, heavy subunit	–
P62424	0.020	0.0016	0.082	1.91E-10	60S ribosomal protein L7a	–
Q4SUA7	0.068	0.0054	0.079	1.21E-39	Prothrombin	F
P62917	0.010	0.00081	0.079	2.85E-05	60S ribosomal protein L8	–
P15880	0.018	0.0013	0.072	2.19E-10	40S ribosomal protein S2	–
P32969	0.0083	0.00054	0.065	0.00022	60S ribosomal protein L9	–
P62841	0.0044	0.00027	0.061	0.044	40S ribosomal protein S15	–
P50914	0.0098	0.00054	0.055	1.47E-05	60S ribosomal protein L14	–
Q93088	0.0054	0.00027	0.050	0.0079	Betaine-homocysteine S-methyltransferase 1	–
UPI0000F213D7	0.036	0.00081	0.022	3.17E-27	Unknown	–
P23396	0.0078	0	0.010	1.04E-05	40S ribosomal protein S3	–
UPI000066106F	0.0083	0	0.010	3.89E-06	Alpha-1-antiproteinase	–
Q4T526	0.003	0	0.010	0.044	Microfibril-associated glycoprotein 4	–
P62888	0.0034	0	0.010	0.044	60S ribosomal protein L30	–
Q4SXM5	0.0059	0	0.010	0.00047	Complement C5 precursor	–
Q8JJ67	0.0039	0	0.010	0.020	Leukocyte cell-derived chemotaxin 2	–
Q4RFG2	0.0034	0	0.010	0.044	Phospholemman	–
Q58FF7	0.0034	0	0.010	0.044	Heat shock protein HSP 90-beta	F
Q4T2B6	0.0049	0	0.010	0.0031	Complement C1q tumor necrosis factor-related protein 3	–
P27635	0.0034	0	0.010	0.044	60S ribosomal protein L10	–

RTCLF_lib1 is the 'Relative Tag Coverage Level of a protein-coding gene Family' of the *N. Angustata* liver transcriptome, and RTCLF_lib2 is that of *D. mawsoni*. The last column 'Agreement' shows the agreement of this comparison to the previous meta-analysis. The letter 'T' indicates a gene family was significant in current analysis and was up- or down-regulated in the same way in the previous meta-analyses; while '–' indicates significant in this analysis but insignificant in the previous analysis; and 'F' in opposing directions between the two comparisons.

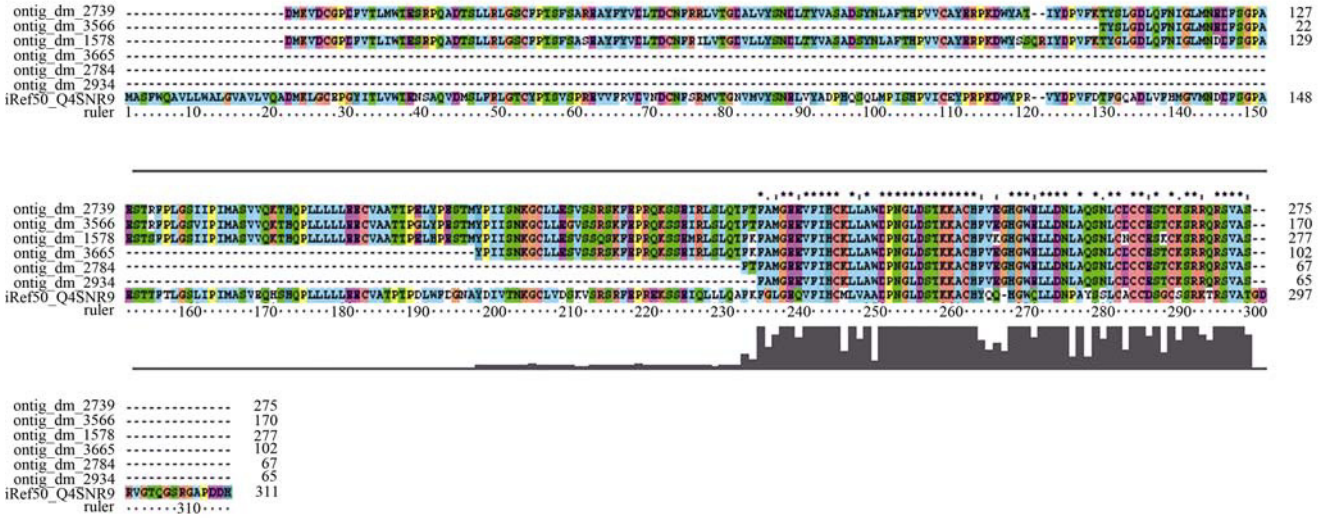

Fig. 4. Multiple sequence alignment of 6 ZPC contigs. ZPC alignment shows that the contigs share high amino acid sequence identity, but were in different lengths resulted from partial sequencing. In such a case, it is more appropriate and convenient to evaluate the transcription levels of all the contigs in a 'family' group.

To further evaluate the validity of the method presented here, we compared results obtained by Pro-Diff and GO-Diff using the same EST data from *D. mawsoni* and *N. angustata*. The result generated by Pro-Diff is a list of differentially expressed gene families while that generated by GO-Diff is a list of differentially represented GO terms. Since the nature of the results from the two methods is different, we must first transform the Pro-Diff result into a GO result using GO over-representation (GOR) analysis (Khatri and Draghici, 2005) before making comparisons.

1. GOR analysis of the Pro-Diff result

GOR analysis with the Pro-Diff results determined whether the particular molecular functions/biological processes/cellular components are over- or under-represented in the subset of differentially expressed gene families within the total representative proteins. Sixty-two GO terms were found to be over- or under-represented. According to the true path rule, daughter nodes in the 62 GO term set were removed, which resulted in 12 non-redundant over- or under-represented GO terms (Table 3).

2. GO-Diff analysis of D. mawsoni and N. angustata transcriptomes

GO-Diff was used to compare the *D. mawsoni* and *N. angustata* transcriptomes directly with the same data of the Pro-Diff-GOR analysis. GO-Diff predicted 623 GO terms as over- or under-represented, and after removal of redundant nodes, leaving 205 GO terms.

3. Comparing GOR and GO-Diff results

Our goal was to determine the extent to which the 12 GO terms discovered by the two-step analysis were included in the 205 GO-Diff results. Of the 12 GO nodes compared with the 205 GO-Diff result nodes, 4 were the same without daughter nodes, 7 were parents of the GO-Diff GO terms, and 1 different (Table 3). Considering the total number of GO nodes of the GO structure was more than 24,000, the two analyses showed high consistency (p-value < 9.7E-6 by a rough estimation using Fisher's Exact test).

Discussion

Considerations of the GO-wise, orthorlog-wise and family-wise comparison methods

The three methods have different features in application. GO-based methods, taking the advantage of a predefined functional structure, identify differentially represented functional units directly. However, as the relationship between genes and GO functional terms is many-to-many, it is hard to discern what genes are of significance, and false positive GO categories will arise when a differentially

Table 3
GO over-representation analysis of Pro-Diff results

GO ID	GOR ratio	p-value	GO term	Name space	Consistency
GO:0003735	8.50	6.3E-07	Structural constituent of ribosome	F	T
GO:0004866	9.96	0.022	Endopeptidase inhibitor activity	F	T
GO:0005615	7.22	0.0049	Extracellular space	C	F
GO:0005830	10.70	0.019	Cytosolic ribosome (sensu Eukaryota)	C	T
GO:0006412	4.69	8.8E-05	Translation	P	T
GO:0006879	25.69	0.0095	Cellular iron ion homeostasis	P	T
GO:0006956	13.76	0.010	Complement activation	P	T
GO:0007338	16.51	0.021	Single fertilization	P	T
GO:0015935	13.13	0.012	Small ribosomal subunit	C	T
GO:0043231	0.43	0.022	Intracellular membrane-bound organelle	C	T
GO:0045087	13.76	0.010	Innate immune response	P	T
GO:0003735	8.50	6.3E-07	Structural constituent of ribosome	F	T

The 'GOR ratio' is the ratio of relative gene family numbers associated with a certain GO term between differentially expressed families and background families. If the 'GOR ratio' is above 1, this GO term is over-represented in the differentially expressed family set, and vice versa. The 'Name space' column shows the categories of the GO terms: 'F' represents molecular function, 'P' represents biological process and 'C' represents cellular component. The 'Consistency' indicates whether the GOR analysis is consistent with GO-Diff analysis, with 'T' for consistent and 'F' for inconsistent.

expressed gene is linked to multiple GO terms. For orthorlog-based method, as the orthorlogy relationship between two species are not always one to one, meaningful information can be lost when comparing distant species. Its application is usually restricted within closely related species.

It is one of the major aims of this paper to provide a common theoretical framework for cross-species transcritome comparison with the sequencing-based data, to extend our previous GO-Diff method. In this scheme, researchers can use the broad-sense 'family' metrics, for example, mapping transcripts to the Pfam (Finn et al., 2008) and InterPro (Hunter et al., 2008) structural units.

In the family-wise method, the granulity of the 'family' to be chosen can depend on the evolutionary distance between the species being compared and can be adjusted in the mapping procedure with user-defined criteria. For example, researchers can relax the BLAST cutoff to increase the granule size. The theoretical framework of this method can also be extended to non-coding RNA transcripts and to cross-species proteomic comparisons if data become available in the future.

Parameter considerations

In some cases, when the denominator is close to zero, TCRF would be unstable and extremely high. The program allows the users to setup an upper-limit TCRF, e.g., the upper-limit of 100 × difference set in the current program. Since the aim of this method is to find out the statistically meaningful differentially expressed gene families, a ratio being 100 × or 1 million times would make no difference in the statistical sense.

In terms of how to select the parameters in the program, we provide a simple guideline here. There are two types of parameters in our system. One is for statistical tests, which are p-value cutoff and TCRF. Obviously, stringent criteria will give more accurate but fewer results. These parameters can be selected according to the data volume to obtain a balance between accuracy and productivity. The other type is the parameters used to group genes into families, for example, the p-value of the BLAST results. These parameters can be selected based on the evolutionary distances between the species in comparison, which will require multiple adjustments to set in a suitable one.

In conclusion, we developed a method for cross-species transcriptome comparisons for sequencing-based transcriptomes. We developed a theoretical framework that uses gene 'families' as units for cross-species comparison. The program we designed used a gene cluster-wise approach as an approximation to the stringent concept of the gene family, and which can be extended to encapsulate other 'family' schemes, such as the Pfam protein family. We provided the program, Pro-Diff for cross-species tran-

scriptome comparisons using this method. Application of this method to a comparison of gene expression in the livers of the related Antarctic *D. mawsoni* and non-Antarctic *N. angustata* showed that this method generated good agreement with biological expectations and with our previous GO-Diff methodology, thus justified the method and confirmed its usefulness.

Acknowledgements

The work was supported by the grants from the Ministry of Science and Technology of China (No. 2006AA02Z331 and 2004CB117404), the Key Project of Chinese Academy of Sciences (No. KSCX2-YW-N-020) to Liangbiao Chen, and NSF OPP 0636696 to C-H CC.

References

Abele, D., and Puntarulo, S. (2004). Formation of reactive species and induction of antioxidant defence systems in polar and temperate marine invertebrates and fish. Comp. Biochem. Physiol. A Mol. Integr. Physiol. **138:** 405–415.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25:** 25–29.

Audic, S., and Claverie, J.M. (1997). The significance of digital gene expression profiles. Genome Res. **7:** 986–995.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B **57:** 289–300.

Binkert, C., Demetriou, M., Sukhu, B., Szweras, M., Tenenbaum, H.C., and Dennis, J.W. (1999). Regulation of osteogenesis by fetuin. J. Biol. Chem. **274:** 28514–28520.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST—database for "expressed sequence tags". Nat. Genet. **4:** 332–333.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. **18:** 630–634.

Caceres, M., Lachuer, J., Zapala, M.A., Redmond, J.C., Kudo, L., Geschwind, D.H., Lockhart, D.J., Preuss, T.M., and Barlow, C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. Proc. Natl. Acad. Sci. USA **100:** 13030–13035.

Chen, J., Blackwell, T.W., Fermin, D., Menon, R., Chen, Y., Gao, J., Lee, A.W., and States, D.J. (2007). Evolutionary-conserved gene expression response profiles across mammalian tissues. Omics **11:** 96–115.

Chen, Z., Wang, W., Ling, X.B., Liu, J.J., and Chen, L. (2006). GO-Diff: mining functional differentiation between EST-based transcriptomes. BMC Bioinformatics **7:** 72.

Chen, Z., Xue, C., Zhu, S., Zhou, F., Ling, X.B., Liu, G.P., and Chen, L. (2005). GoPipe: streamlined gene ontology annotation for batch anonymous sequences with statistics. Prog. Biochem. Biophy. **32:** 187–191.

Chen, Z., Cheng, C.H., Zhang, J., Cao, L., Chen, L., Zhou, L., Jin, Y., Ye, H., Deng, C., Dai, Z., Xu, Q., Hu, P., Sun, S., Shen, Y., and Chen, L. (2008). Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. USA **105:** 12944–12949.

Clarke, A., Doherty, N., DeVries, A.L., and Eastman, J.T. (1984). Lipid content and composition of three species of Antarctic fish in relation to buoyancy. Polar Biol. **3:** 77–83.

Eastman, J.T., and DeVries, A.L. (1981). Buoyancy adaptations in a swim-bladderless Antarctic fish. J. Morph. **167:** 91–102.

Eastman, J.T., and DeVries, A.L. (1982). Buoyancy studies of notothenioid fishes in McMurdo Sound, Antarctica. Copeia **2:** 385–393.

Enard, W., Khaitovich, P., Klose, J., Zollner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G.M., Bontrop, R.E., and Paabo, S. (2002). Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340–343.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:** 186–194.

Fei, Z., Tang, X., Alba, R.M., White, J.A., Ronning, C.M., Martin, G.B., Tanksley, S.D., and Giovannoni, J.J. (2004). Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. Plant J. **40:** 47–59.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., and Bateman, A. (2008). The Pfam protein families database. Nucleic Acids Res. **36:** D281–D288.

Gilad, Y., Rifkin, S.A., Bertone, P., Gerstein, M., and White, K.P. (2005). Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Res. **15:** 674–680.

Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P., and White, K.P. (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature **440:** 242–245.

Gu, X., and Su, Z. (2007). Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. Proc. Natl. Acad. Sci. USA **104:** 2779–2784.

Heiss, A., DuChesne, A., Denecke, B., Grotzinger, J., Yamamoto, K., Renne, T., and Jahnen-Dechent, W. (2003). Structural basis of calcification inhibition by alpha 2-HS glycoprotein/fetuin-A. Formation of colloidal calciprotein particles. J. Biol. Chem. **278:** 13333–13341.

Hoffmann, R., Lottaz, C., Kuhne, T., Rolink, A., and Melchers, F.

(2007). Neutrality, compensation, and negative selection during evolution of B-cell development transcriptomes. Mol. Biol. Evol. **24:** 2610–2618.

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. Genome Res. **9:** 868–877.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H., and Yeats, C. (2008). InterPro: the integrative protein signature database. Nucleic Acids Res. **21:** 21.

Khaitovich, P., Paabo, S., and Weiss, G. (2005a). Toward a neutral evolutionary model of gene expression. Genetics **170:** 929–939.

Khaitovich, P., Enard, W., Lachmann, M., and Paabo, S. (2006). Evolution of primate gene expression. Nat. Rev. Genet. **7:** 693–702.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., and Paabo, S. (2004). A neutral model of transcriptome evolution. PLoS Biol. **2:** E132.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M., and Paabo, S. (2005b). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309:** 1850–1854.

Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics **21:** 3587–3595.

Lemos, B., Bettencourt, B.R., Meiklejohn, C.D., and Hartl, D.L. (2005). Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol. Biol. Evol. **22:** 1345–1354.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. (2000). An optimized protocol for analysis of EST sequences. Nucleic Acids Res. **28:** 3657–3665.

Liao, B.Y., and Zhang, J. (2006a). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol. **23:** 530–540.

Liao, B.Y., and Zhang, J. (2006b). Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol. Biol. Evol. **23:** 1119–1128.

Man, M.Z., Wang, X., and Wang, Y. (2000). POWER_SAGE: comparing statistical tests for SAGE experiments. Bioinformatics **16:** 953–959.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:** 376–380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. **18:** 1509–1517.

Metcalf, V.J., Brennan, S.O., and George, P.M. (1999). The Antarctic toothfish (*Dissostichus mawsoni*) lacks plasma albumin and utilises high density lipoprotein as its major palmitate binding protein. Comp. Biochem. Physiol. B Biochem. Mol. Biol. **124:** 147–155.

Meyers, B.C., Vu, T.H., Tej, S.S., Ghazal, H., Matvienko, M., Agrawal, V., Ning, J., and Haudenschild, C.D. (2004). Analysis of the transcriptional complexity of *Arabidopsis* thaliana by massively parallel signature sequencing. Nat. Biotechnol. **22:** 1006–1011.

Nobuta, K., Vemaraju, K., and Meyers, B.C. (2007). Methods for analysis of gene expression in plants using MPSS. Methods Mol. Biol. **406:** 387–408.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., Gonzalez, J.R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W., and Hurles, M.E. (2006). Global variation in copy number in the human genome. Nature **444:** 444–454.

Sutton, G., White, O., Adams, M., and Kerlavage, A. (1995). TIGR Assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci. Tech. **1:** 9–19.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23:** 1282–1288.

t Hoen, P.A., Ariyurek, Y., Thygesen, H.H., Vreugdenhil, E., Vossen, R.H., de Menezes, R.X., Boer, J.M., van Ommen, G.J., and den Dunnen, J.T. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. **36:** e141.

Tirosh, I., and Barkai, N. (2008). Evolution of gene sequence and gene expression are not correlated in yeast. Trends Genet. **24:** 109-113.

van Ruissen, F., and Baas, F. (2007). Serial analysis of gene expression (SAGE). Methods Mol. Biol. **383:** 41–66.

Whitehead, A., and Crawford, D.L. (2006). Neutral and adaptive variation in gene expression. Proc. Natl. Acad. Sci. USA **103:** 5425–5430.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. **34:** D187–D191.