

# Gene2Path: A Data Analysis Tool to Study Fish Gene Pathways by Automatic Search of Orthologous Genes

Natalia Ballesteros<sup>1</sup>, Néstor Aguirre<sup>2</sup>, Julio Coll<sup>3</sup>, Sara I Pérez-Prieto<sup>1</sup> and Sylvia Rodríguez Saint-Jean<sup>1\*</sup>

<sup>1</sup>Centro de Investigaciones Biológicas (CSIC), C/ Ramiro de Maeztu 9, 28040 Madrid, Spain

<sup>2</sup>Instituto de Física Fundamental (CSIC), C/ Serrano 123, 28006 Madrid (Spain)

<sup>3</sup>Instituto Nacional de Investigaciones Agrarias (INIA), Ctra La Coruña km7, Madrid 28040, Spain

## Abstract

Most of the gene regulation pathways data from biochemical and molecular experiments are drawn from humans or from species commonly used as experimental animal models. Accordingly, the software packages to analyse these data on the basis of specific gene identification codes (IDs) or accession numbers (AN) are not easy to apply to other organisms that are less characterized at the genomic level. Here, we have developed the Gene2Path programme which automatically searches pathway databases to analyse microarray data in an independent, species-specific way. We have illustrated the method with data obtained from an immune targeted rainbow trout microarray to search for orthologous pathways defined for other well known biological species, such as zebrafish, although the software can be applied to any other case or species of interest. The scripts and programme are available and free at the "GENE2PATH" web site <http://gene2path.no-ip.org/cgi-bin/gene2path/index.cgi>. A user guide and examples are provided with the package. The Gene2Path software allows the automated searching of NCBI databases and the straightforward visualization of the data retrieved based on a graphic network environment.

**Keywords:** Pathway analysis tool; Microarrays; Species-independent pathway analysis; Orthologous genes

## Introduction

The use of high-density microarrays had a significant impact on studies of gene expression, attracting much interest among biologists. Microarray technology have been used to test the expression of thousands of genes in a single experiment, exploiting the ability of messenger RNA (mRNA) to bind specifically to the DNA template from which it was derived. Microarray gene expression screening can identify the genes involved in a given process, as well as predict interactions among thousands of genes by studying genome transcription. Many fields have benefited considerably from DNA microarray technology, such as drug discovery and toxicological research [1,2], as well as human disease diagnosis. However, studies in the field of veterinary sciences have been more restricted due to the lack of their genome sequences. In fish, for instance, most microarray and/or RNAseq studies have been carried out on zebrafish (*Danio rerio*), which is a well characterized species with large amount of sequenced genome for which commercial arrays are available. However, other economically relevant cultured fish are still far from having their complete genome sequenced, such as turbot (*Scophthalmus maximus*), sea bass (*Dicentrarchus labrax*), sea bream (*Sparus aurata*). Although some commercial microarrays have recently been made available for some of these species such Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss* Walbaum); their genome coverage is still far from that of the zebrafish [3].

Several studies have highlighted the importance of presenting microarray data in the framework of documented biological pathways [4,5]. Typically, microarray gene expression experiments produce long lists of genes that are differentially expressed in two different circumstances. The information regarding these pathways is difficult to apply to species that are not well characterized, mainly because most of the actual software packages use species-specific gene identification data (IDs) from a few biological species that cannot handle genomic data for other less well known species.

To circumvent this problem it has been proposed that "...most of the genetics and physiology of the less well-represented species will be

similar or comparable with the data of human and laboratory animal species stored in the database" [6]. Thus, ranking pathways in terms of their relevance to a particular phenotype or metabolic route can help researchers focus on a few sets of genes and such an approach may be very useful to answer some biological hypotheses.

However most biologists and veterinarians who are not familiar with simultaneous upload of thousands of data may have some difficulty because both, the need to configure on a local computer and the excessively long computing times required for analyzing several genes at once, are prohibitive.

In this study, we present an open access web server called Gene2Path for analyzing microarrays results and automatic searches of orthologous genes to be associated in pathways.

Therefore, the main objective of the present study was to develop and validate a tool that extrapolates information associated with different pathways across different species. The programme provides a software tool that uses species-independent gene IDs and streamlines that process searching for information regarding pathways in online public databases. The software uses lists of genes in microarrays (IDs), combining pathway information with this microarray data. Other researchers working with less well studied species, such as the chicken, have also faced the same problem when analyzing the pathways associated with the data derived from microarrays. A tool to study a

**\*Corresponding author:** Sylvia Rodríguez Saint-Jean, Centro de Investigaciones Biológicas (CSIC), C/ Ramiro de Maeztu 9, 28040 Madrid, Spain, Tel: 34918373112; Fax: 34915360432; E-mail: [sylvia@cib.csic.es](mailto:sylvia@cib.csic.es)

**Received** February 10, 2015; **Accepted** February 26, 2015; **Published** March 20, 2015

**Citation:** Ballesteros N, Aguirre N, Coll J, Pérez-Prieto SI, Saint-Jean SR (2015) Gene2Path: A Data Analysis Tool to Study Fish Gene Pathways by Automatic Search of Orthologous Genes. J Aquac Res Development 6: 329. doi:[10.4172/2155-9546.1000329](https://doi.org/10.4172/2155-9546.1000329)

**Copyright:** © 2015 Ballesteros N, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

chicken-specific reaction to bacterial infection was developed by Pas et al. [6] based on a set of PERL scripts to extract data from databases via internet and using the names of the genes in the microarray. The gene short names and synonyms were extracted from the GO and KEGG databases, and the results were visualized using colour codes when they combined the pathway data with the microarray data.

In our case example rainbow trout-specific pathways are not available. Therefore, we used pathway schemes from other species (such as zebrafish), to obtain orthologous pathways. There are some free programs for the analysis of data from microarrays in an specific pathway mammal, but they mostly work for human, mouse or other models species not for commercial and marine fish. Our Gene2Path program has the so called "Orthology step", which offers the possibility to detect homology between the non annotated genes or sequences of the uncommon specie with other eukaryote genomes and annotated gene IDs.

We are interested in teleost fish virology and immunology, and accordingly we validated this tool by identifying immune and infection-related signalling pathways for fish species and chose for that the KEGG database (<http://www.genome.ad.jp/kegg/>).

Infectious pathogens are a serious problem in aquaculture, and salmonid fish viruses are responsible for important losses in rainbow trout and salmon farming, reflected in the intense research into these viruses within the field of aquaculture [7-9]. Infectious pancreatic necrosis virus (IPNV) for example is the aetiological agent of a well characterized acute disease that produces systemic infection and relevant mortality in farmed rainbow trout (*Oncorhynchus mykiss*) and other salmonid species. The mortality of this virus may be as high as 70% in young fish. The virus establishes an asymptomatic carrier state in survivors [10], both in different species of salmonids and in other species of farmed fish such as turbot and Atlantic cod (*Gadus morua*). Nevertheless, the production of vaccines against this virus is an area that has been little investigated.

In a previous study, we assayed oral DNA-based immunotherapy against IPNV [11] and the immune specific host reaction to a VP2-IPNV vaccine). The transcriptional changes produced by infection were determined in a rainbow trout 15k microarray designed by including annotated genes selected by key-words in the GenBank. However, since difficulties arose when trying to analyze the results in terms of pathways, we tried to solve those problems by designing a user friendly and amenable programme to study this data, which is described below. We have illustrated the method with data obtained from this rainbow trout microarray to search for orthologous pathways in zebrafish, although the software can be applied to any other case or species of interest. Finally, the present work reports the search of some orthologous genes or proteins involved in several pathways from three teleost fish species (*Dicentrarchus labrax*, *Salmo salar* and *Oncorhynchus mykiss*) from KEGG database.

## Materials and Methods

### Database searches

We have used the following databases to design the programme's algorithm: (See 132 Supplementary File1) Nucleotide and Expressed Survey Sequence (EST): <http://www.ncbi.nlm.nih.gov/nuccore/> and <http://www.ncbi.nlm.nih.gov/nucest/>; The Nucleotide Genome Survey Sequence (GSS) and Expressed Sequence Tag (EST) databases contain nucleic acid sequences typically uncharacterized such as short genomic (GSS) or cDNA (EST) sequences.

**HomoloGene:** (<http://www.ncbi.nlm.nih.gov/homologene/>).

It is a program that makes use of amino acid sequence searches (blastp) to find more distant relationships, although the procedure still refers to the DNA sequence to perform some of the statistics. Moreover, HomoloGene entries now include paralogues in addition to orthologues. Nevertheless, data for all species is still not available. For example, fish are only represented by zebrafish (*Danio rerio*).

**Gene** (<http://www.ncbi.nlm.nih.gov/gene/>): This database contains information on gene specificity, structure, function, homology between species and citations. The database supplies gene-specific connections in the nexus of a map, sequence, expression, structure, function, citation and homology data.

**UniGene** (<http://www.ncbi.nlm.nih.gov/unigene/>): This database groups transcript sequences from different loci based on genomic sequences. The availability of a genomic sequence is helpful to identify sets of transcript sequences that correspond to distinct transcript loci or to annotated genes.

**Blast2GO** (<http://www.blast2go.com/b2ghome>) is a programme to get homologous amino-acid sequences from nucleotide sequences. This research tool uses BLASTx to find the most similar sequence between several input sequences in a FASTA format.

**KEGG** (<http://www.genome.ad.jp/kegg/>) is a collection of manually drawn pathway maps, representing the molecular interaction and reaction networks for a number of cellular processes and genetic events. The database contains gene names and information on biological species-specific pathways. While searching the KEGG database with known pathways, we found that genes may be represented with several synonyms not all of which were linked to a pathway. Therefore, when the species of interest corresponding ID was not available in the KEGG database, it was first necessary to find the gene ID of the corresponding orthologous gene. Our programme provides this tool.

### Microarray data

We previously studied the transcriptional changes induced by an oral DNA vaccine against the IPNV by using a rainbow trout microarray. Pooled RNA from the fish was hybridized on an Agilent rainbow trout microarray (8x15K format custom microarrays -ID032303-) containing 6442 60-mer oligo sequences. The annotation file of the microarrays was designed by us and might be provided by the supplier in ".txt" format (Ballesteros et al., 2012 for further details of the microarray used, the hybridization conditions and the first analysis). For the corresponding raw data see NCBI (<http://www.ncbi.nlm.nih.gov/geo/>), accession Number GSE31591. We used this microarray data as an example for the analysis and interpretation of the results from the Gene2Path programme.

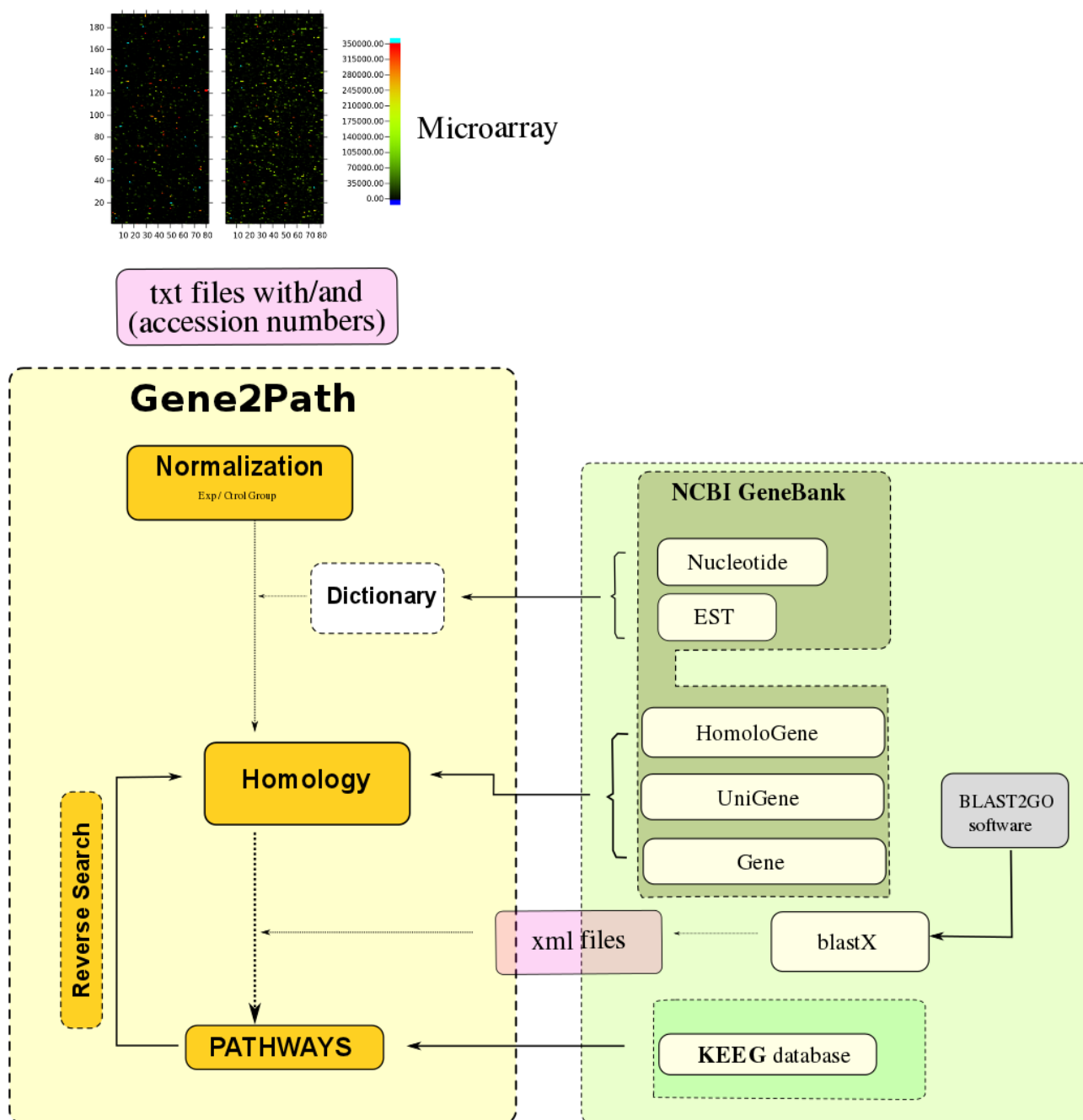
## Results

### Software package for the analysis and interpretation of DNA microarray data

An scheme of the steps to follow the Gene2Path programme is shown in Figure 1, briefly the next steps are shown in Figures 2 and 3 (Supplementary File 2):

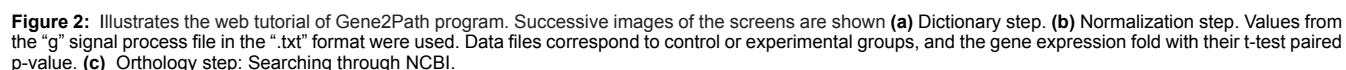
### Dictionary, gene symbol and download of the FASTA data file

The dictionary section of the programme Gene2Path converts an input as accession numbers IDs into other code types available in the NCBI database (such as gene identification -GI- in GenBank).




**Figure 1:** Scheme of the pathway analysis program and its search into different databases such as NCBI and KEGG. The diagram of Gene2Path is represented in the yellow cells, showing the automated steps of the program. The cells in green represent the databases or external programs used by Gene2Path. Arrows indicate the databases needed to follow up each one of the program steps. Into the rose cells are mentioned the archives extensions used by Gene2Path: txt are the gProcessedSignal of each of the genes from the microarray "input data"; the xml archives were obtained from the external program BLAST2GO and are used by the Gene2Path program to process and organize the blastX information. The Dictionary section provided, if needed, the gene identification codes (ID), obtained from the NCBI GeneBank for each one of the genes.

The scripts and programme are available at the "GENE2PATH" web site and they include the following features: (1) The addition of synonyms for gene identification by searching the NCBI database; (2) the analysis of the intensity of the gProcessedSignal reading of the synonym names microarray input files (extension txt); (3) Automated gene homology searches through the Homologene, Gene and Unigene NCBI databases; (4) An automatic filter was developed to find sequences similar by BLAST and to visualize the xml output file and to deal with several files at the same time; (5) Search for pathway information in the KEGG database using Gene IDs.



a.

**Centro de Investigaciones Biológicas**



Department of Molecular Microbiology  
and Infection Biology

**Vaccines and Gene Expression Group**

**Gene2Path: Analyze microarray data  
and pathway from different species**

**Home**

**Tutorial**

**Programs**

Dictionary

Microarray analysis

Homology

Pathways

Reverse Search

## Homology

Search type: blast2go

**Enter accession numbers**

☐ **Species(s) list**

**Species**

**Enter accession numbers**

**Species**

blast2go file (\*.xml)


Choose File blastResult.xml ☐ **Species(s) list**

**Species**

Send

**b.**

**Centro de Investigaciones Biológicas**



Department of Molecular Microbiology  
and Infectious Biology

**Vaccines and Gene Expression Group**

**Gene2Path: Analyze microarray data  
and pathway from different species**

**Home**

**Tutorial**

**Programs**

Dictionary

Microarray analysis

Homology

Pathways

Inverse Search

**Contact**

### KEGG Pathways

Search Type: Collective ▾

Enter Gene ID	Enter Gene IDs
<p><input type="text"/></p> <p>Species (codes <a href="#">help</a>)</p> <p><input type="text"/></p> <p><input type="checkbox"/> Live</p> <p><input type="checkbox"/> Live and details</p>	<p>572053 262557</p> <div style="border: 1px solid #ccc; height: 150px; margin-top: 5px;"></div> <p>Species (codes <a href="#">help</a>)</p> <p><input type="text"/></p>

Send

**C.**

Centro de Investigaciones Biológicas

Department of Molecular Microbiology  
and Infection Biology

Vaccines and Gene Expression Group

Gene2Path: Analyze microarray data  
and pathway from different species

Home

Tutorial

Programs

Dictionary

Microarray analysis

Homology

Pathways

Inverse Search

Contact

FEQO ID

Species (codes [here](#))

☒ To obtain FASTA sequence(s)

# Centro de Investigaciones Biológicas

Department of Molecular Microbiology  
and Infection Biology

Vaccines and Gene Expression Group

Gen2Path: Analyze microarray data  
and pathway from different species

Home

Tutorial


Programs

Dictionary  
Microarray analysis  
Homology  
Pathways  
Inverse Search

Contact

Gen2Dbase	Gen2D	Gen2Syn	Gen2Path	preX2Ch	kb	Prot Description
21087938	---	A2421440.1	AD213109	77.0	0.0	CCR (chemokine 4 [Cytosines caspase])
21087918	---	AB014657.1	BRAB1450	61.0	0.0	CCR (chemokine receptor-1 [Cytosines caspase])
20857100	---	U0161841.1	AD101040	9.0	0.0	interferon-3 [Cytosines caspase])
56743197	---	U0301010.1	AD135377	48.0	0.0	growth hormone receptor type 2b [Cytosines caspase]
57974696	---	U0301010.1	AD135377	48.0	0.0	growth hormone receptor type 2b [Cytosines caspase]
21087921	---	A2421440.1	AD213109	77.0	0.0	CCR (chemokine 4 [Cytosines caspase])
44502196	---	AB211697.1	AD244111	77.0	0.0	carboxin B peptidoglycan [Cytosines caspase]
48796420	---	A2530424.1	AD793312	26.0	0.0	CCR (chemokine 4 protein [Cytosines caspase])
21087928	---	A2421440.1	AD213109	77.0	0.0	CCR (chemokine 4 [Cytosines caspase])
46767690	---	U0321958	AEAD7020	64.6	0.0	interleukin 12a [Cytosines caspase]
103212977	---	A2421440.1	AD213109	77.0	0.0	CCR (chemokine 4 [Cytosines caspase])
103212950	---	D0400124.1	AD015950	26.7	0.0	CD4 [Cytosines caspase]
230211600	---	AB201597.1	CAE51212	44.0	0.0	IL12: leucocytes receptors [Cytosines caspase]
77454500	---	A2530424.1	AD793312	26.0	0.0	CCR (chemokine 4 protein [Cytosines caspase])
56743200	---	A2621155.1	CAG45702	55.2	0.0	interleukin-21 [Cytosines caspase]
21097921	---	AB014657.1	BRAB1450	60.0	0.0	CCR (chemokine receptor-2 [Cytosines caspase])
79098972	---	U0161841.1	AD101040	9.0	0.0	interferon-3 [Cytosines caspase]
208597114	---	U0161841.1	AD101040	63.6	0.0	interferon-3 [Cytosines caspase]

**Centro de Investigaciones Biológicas**



Department of Molecular Microbiology  
and Infection Biology

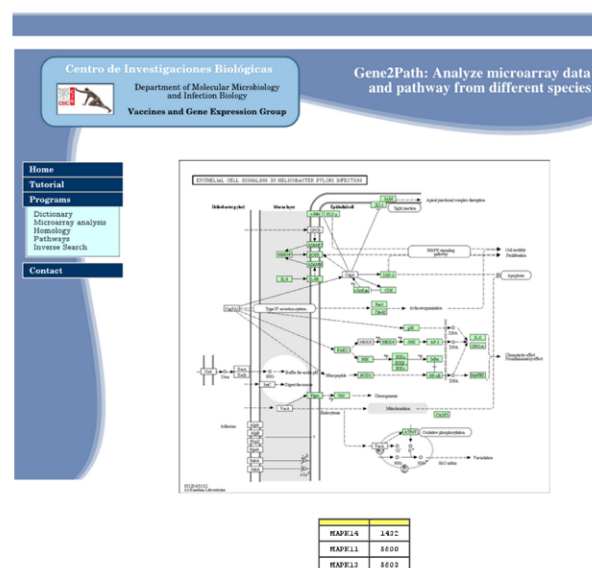
**Vaccines and Gene Expression Group**

## Gene2Path: Analyze microarray data and pathway from different species

- [Home](#)
- [Tutorial](#)
- [Programs](#)
  - [Dictionary](#)
  - [Microarray analysis](#)
  - [Homology](#)
  - [Pathways](#)
  - [Inverse Search](#)
- [Contact](#)

### KEGG Pathways

<a href="#">Axi2D161</a>	292157
<a href="#">Axi2D162</a>	872655
<a href="#">Axi2D163</a>	292157
<a href="#">Axi2D164</a>	292157
<a href="#">Axi2D165</a>	292157



**Figure 3:** Illustrates the web tutorial of Gene2Path program. Successive images of the screens are shown (a) Orthology step: Searching through Blast2GO program. (b) Identifying pathways through KEGG. Gene IDs are necessary to search genes in the KEGG pathways involved. (c) Reverse Search of pathway. The web page shows the boxes or components separated by horizontal broken lines, which permits the genes to be assigned to a particular box.



Moreover, it is possible to obtain the corresponding gene sequences in FASTA format and to visualize the gene symbol (short name) for each of the genes in the microarray. The IDs given by the user are run through the Nucleotide or EST sequence database, named “nucore” or “nucust”, respectively (Figure 2a) and a final output list is provided.

### Normalization and the level of gene expression from DNA microarray data

The normalization section of programme Gene2Path automates the analysis of the data generated by the microarray scanner. In our case example, values from the “g” signal process file in “.txt” format were used and most simple calculations were chosen (Supplementary File 3). Among the programme’s options, the user can choose whether the data files correspond to control or experimental groups, and the level of gene expression to be visualized on the web site <http://gene2path.no-ip.org/cgi-bin/index.cgi> can be established. In our example, values  $\geq 2$  fold were selected. Finally, the program facilitates the process of data with t-test to obtain the p-value for statistics (Figure 2b).

## Orthology

### Identifying gene orthologous through NCBI

The programme allows the search of similarity of a deduced amino acid sequence from translated nucleotide to be determined between two species selected by the user. The procedure involves first searching for gene IDs in the GenBank web using Unigen, Gene and HomoloGene to detect orthology between the annotated genes or sequences of entire eukaryote genomes (Figure 2c). The results are visualized on the web and the percentage of homology is shown.

### Identifying orthologous genes through BLAST2GO

Gene2Path filters and organizes the results from the Blast2GO program (“.xml” files). Blast2GO uses the deduced amino acid sequence in order to find orthologous proteins in other species (blastx). As FASTA sequences from each of the genes are needed to run the Blast2GO programme, we can use data obtained in the Dictionary section (Figure 3a). The programme produces a table with information such as: The gene ID from the original or input species, gene ID in the orthologous selected by the user, gene symbol (short name) of the orthologous gene, GenBank entry of the gene sequence from the orthologous sequence, protein accession ID of the orthogous protein, percentage of homology and a short description of the gene (Figure 3a, right side).

### Identifying KEGG pathways for the orthologous genes

The Gene2Path programme finds routes defined by the genes (pathways) in the KEGG database. Gene IDs are necessary to search for KEGG pathways containing those genes and gene IDs can be obtained from the orthologous genes, as described above. Each pathway is identified by its own ID (KEGG alias). To automate the search and the retrieval of the gene data from the KEGG database, a BASH script was written using the KEGG API. Direct links in each pathway were added to the file for each of the genes. This software can be used freely <http://gene2path.no-ip.org/cgi-bin/index.cgi>, with no need to register (Figure 3b).

### Reverse Search of Pathway

The programme provides a tool to find genes involved in specific pathways by using the KEGG ID and recovering each of the genes shown in the pathway box through their gene IDs. The web page shows the boxes or components separated by horizontal broken lines,

which permits the genes to be assigned to a particular box (Figure 3c). The user may obtain the sequences for each of the genes in FASTA format for the procedure. An example of a selected pathway and table, with the names of the boxes or KEGG components, as well as their corresponding Gene IDs is shown. This tool can be used to search for gene or protein orthologous involved in pathways because not all the pathways are available for a particular species in the KEGG data base. In this study, we selected potentially-relevant zebra fish and human (*Homo sapiens*, hsa and *Danio rerio*, dre respectively) pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (<http://www.genome.ad.jp/kegg/>) for *Dicentrarchus labrax*, *Oncorhynchus mykiss* and *Salmo salar*.

The human (“hsa”) and zebra fish (“dre”) pathways were selected because they are the most complete and phylogenetically close, respectively, to our species interest. The KEGG pathways selected for study were: “Mapk signaling-dre04010; Apoptosis- dre04210; TGF-BETA signaling-dre04350; Toll-like receptor-dre04620; NOD-like receptor-dre04321; Cytosolic DNA-sensing-dre04623; Jak-stat signaling-dre04630; Herpes simple Infection-dre05168; Chemokine signaling-hsa04062; B-cell receptor-hsa04662; Fc-epsilon RI signaling-hsa04664; Bacterial invasion-hsa05100; Hepatitis C-hsa05160; Measles virus-hsa05162; Influenza A-hsa05164; HTLV-1-hsa05166 and NK-cell mediated-hsa04650”. Some of these mammalian pathways have unknown fish equivalents. On the other hand, four of seventeen pathways were not found to *D. labrax* (see Supplementary File 4). The Mapk signalling pathway-dre04010 for all species is showed in Figure 4.

In summary, given one gene ID the user of Gene2Path program would be able to obtain which KEGG pathway is in. And given a pathway the user would get how many of the genes are in it.

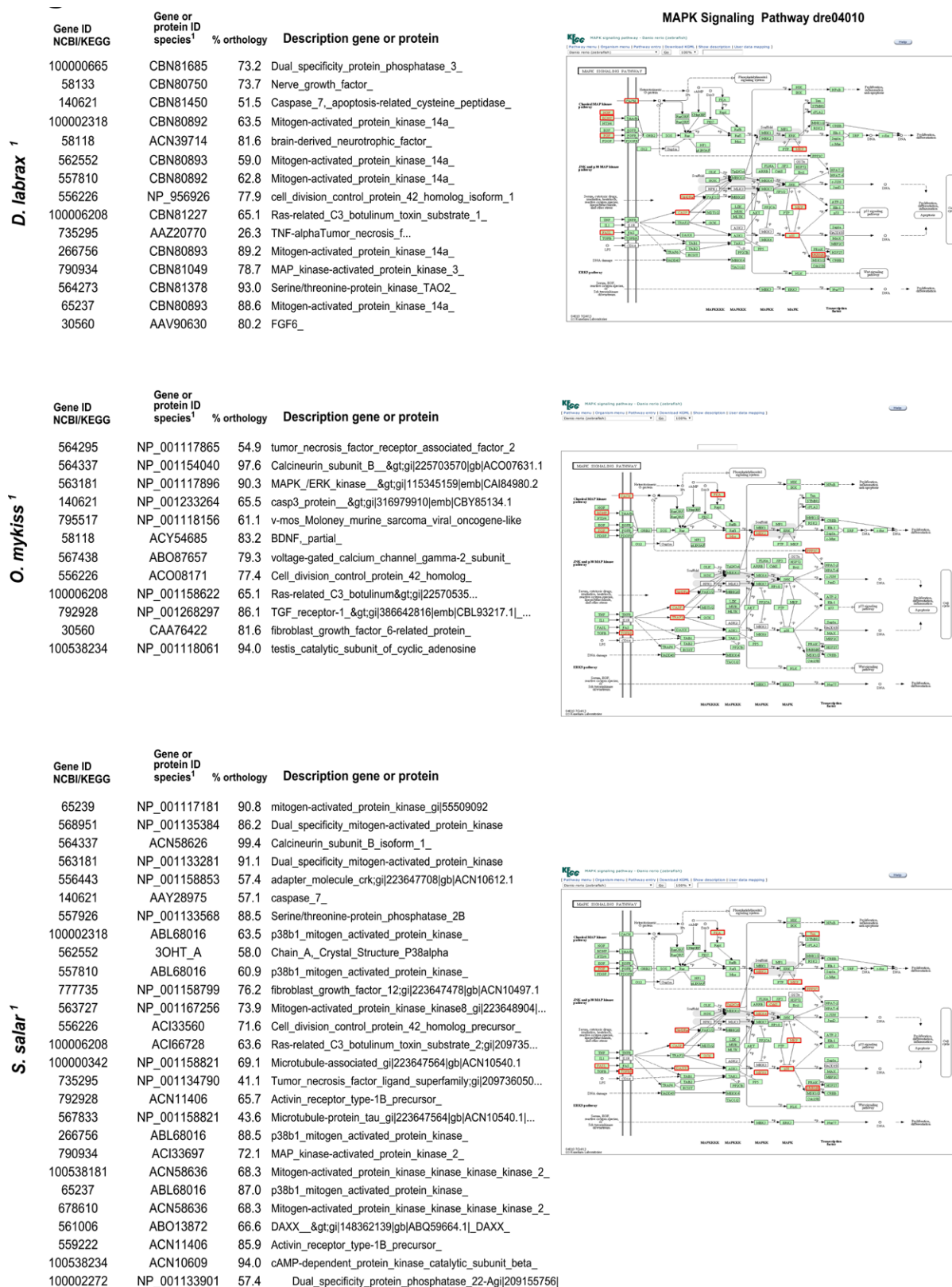
### Gene2Path analysis of the microarray gene expression results of rainbow trout genes following the administration of an oral fish DNA vaccine. A case study

To demonstrate the use of the Gene2Path, we used a dataset derived from an experiment previously reported. Thus, an oral alginate-microencapsulated DNA vaccine against VP2- IPNV protected in rainbow trout *Oncorhynchus mykiss* [11] against IPNV. The kidney and pyloric ceca from vaccinated and control fish, obtained 7 days post vaccination, were assayed using a microarray enriched in rainbow trout immune-related genes from the GeneBank and selected genes from a previous design. Our rainbow trout microarray (8x 15K), called minitroutr 12.8 (Agilent ID032303), contains 6,442 unique 60-mer oligo sequences, each in duplicates arranged randomly in the microarray [11,12].

### Data analysis procedure

**Step 1:** Obtaining the rainbow trout gene IDs. Gene2Path was used to convert the rainbow trout mRNA accession number used for the microarray design into Gene IDs.

**Step 2:** Searching for orthologous genes in zebrafish. (Orthology step). The gene IDs or the accession numbers were used to search homoloGene, Gene and Unigene sections of the NCBI database, and BLASTx using BLAST2GO software for zebrafish (*Danio rerio*) orthologous. This step was obligatory to subsequently search for pathways in the KEGG database, as they are only available for some species. In this example from an input of 6442 rainbow trout accession numbers 1,282 orthologous IDs for zebrafish genes were found in the NCBI database.



**Figure 4:** Results obtained after running the Reverse Search step of Gene2Path program in the MAPK signalling pathway dre04010 of three fish species. Headings are: the gene ID from *D. rerio* (first column), the gene ID of the fish species under study (second column), the percentage orthology (third column) and a short description of the gene or protein of the species selected (fourth column). The figure (right side) illustrates the situation of orthologous genes into the pathway.

**Step 3:** Search zebrafish pathways with orthologous genes: It is interesting to note that only 12 genes into 10 different pathways were detected in KEGGS by using the ID gene symbols provided by the microarray. This search was run without the orthology step (original data <https://earray.chem.agilent.com/earray/search.do?search1/4arrayDesign>); however, after running the Gene2Path program (accession data 2013-01-15), the numbers increased to 1169 genes and 179 pathways (Figure 5). The identification of 169 additional pathways with the Gene2Path programme demonstrates its efficacy. Although genes may be active in more than one pathway. The microarray used was designed to determine transcriptional changes in selected immune genes induced by a DNA vaccine and hence, identification of immune related pathways is to be expected. Nevertheless several other pathways were also detected, such as those involved in “apoptosis and the cell cycle” (23 different KEGG pathways), “regulation of energy metabolism” (42 different KEGG pathways), and several pathways that could be grouped without direct network associations. Figure 6 illustrates the data of the intestinal immune network for IgA production (KEGG alias: dre04672).

Only two genes were obtained manually from the KEGG database, without using the “orthology step” are represented in Figure 6a. The same pathway is shown after running the “orthology step” on the Gene2Path software (Figure 6b); from the 55 genes of the pathway, 18 genes were detected, which represents a 33% increase. IgA production pathway has not been yet identified in fish, but functional similarities with the IgT genes implicated in the *Danio rerio* orthologous could exist, and pathways showed similar trends than most of those pathways described.

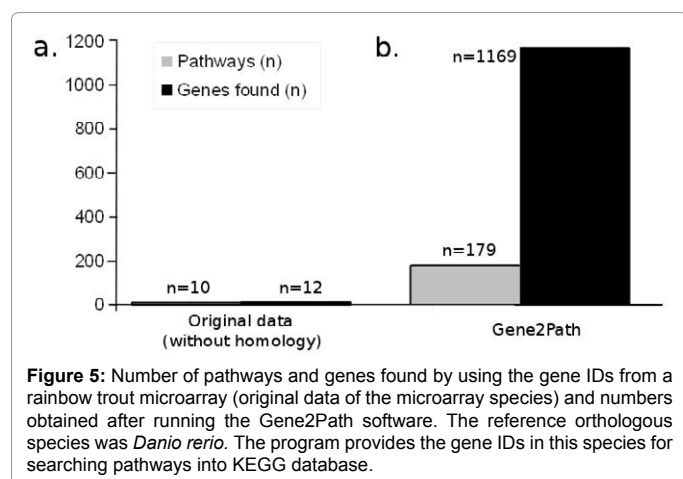
Moreover if a user needs to know the specific genes involved in a pathway, or if there is interest in focusing on a specific section of the pathway, it is possible to obtain these genes through the “Reverse Search of pathways” application of Gene2Path. The results of the pathway analysis of the microarray data provide insight into differential organ-specific biological processes that may explain the differences in host response to the VP2-IPNV vaccine.

## Discussion

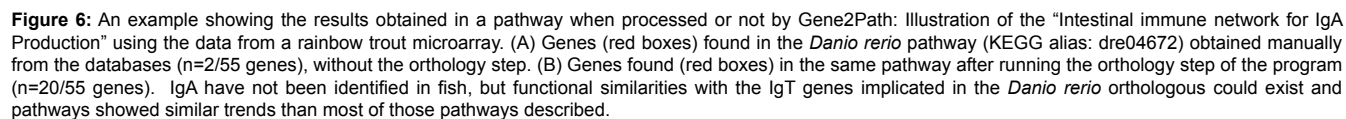
In this article, we describe a pathway-based approach to analyze microarray data from uncommon biological species using newly designed Gene2Path software. To achieve this, we relied on the orthology of the genes identified in a microarray with those from other species included in the KEGG database. The programme

automatically searches pathway databases to analyse microarray data in an independent, species-specific way. (For installation, instructions, examples and source code see Supplementary File 5). We have illustrated the method with data obtained from a rainbow trout microarray to search for orthologous pathways in other well known biological species, such as zebrafish, although the software can be applied to any other case or species of interest. Large scale gene expression studies represent an important advance in experimental molecular biology. Microarrays have become an important tool in functional genomics studies and they are often used to address a variety of biological situations but in some cases they are structured to well know biological species [3]. Such studies rapidly generate large quantities of gene expression data, the handling of which represents a major challenge for biologists. Indeed, the importance of presenting microarray data in the framework of documented biological pathways has often been noted [4,5]. In biology, pathway is a set of interactions or functional relationships between the physical and genetic components of a cell that operate in concert to fulfil a biological requirement. The databases that capture information on these functional interactions of molecular species are numerous, and the lack of uniformity of models and the methods to access this data makes integrating pathway data extremely difficult for uncommon biological species. Thus most of the software packages use species-specific gene IDs and they cannot handle gene data from other species. Yet it is necessary to make such pathway information systems more flexible and efficient, since while data for humans and common laboratory animals such as mice are widely available in databases through the internet, this is not so for other species such as those economically important fish species. The zebrafish (*Danio rerio*) is a model organism for genomic studies and a variety of functional pathways from this species can be found in the databases. The same is true for human databases, the most studied species at the genomic level. These pathways can be used as models to integrate and visualize data from microarray experiments from other species (rainbow trout in our case).

The microarray pathway analysis tool described here can be applied to a typical experiment in which two conditions are compared to identify genes whose differential expression changes significantly with respect to the reference condition. We used a microarray to analyse the differential expression of immune-related genes induced by the administration of an oral DNA vaccine in two rainbow trout organs, a species that is much less well represented than zebrafish in the pathway databases. The principal motivation for building pathway databases is to make tools available that help answer specific biological questions. The majority of genes in most genomes have no known function and examining genes in the context of a particular pathway may help to elucidate their role. For example in our case, a gene of unknown function connected to a set of genes involved in early immunity is likely to also act in this process. However, the power of many pathway analysis techniques is proportional to the amount of input data. Rainbow trout-specific pathways are not available and therefore, we have used pathway data from other species (such as humans or zebrafish), for comparison. Pathway analysis software tools, such as STARNET 2 [13], Reactome database [14] and CYTOSCAPE [15], are available, although again they are only applicable to humans and some experimental animals (mouse and rat). On the other hand, Babelomic (<http://babelomics.bioinfo.cipf.es>), GEPAS (<http://www.gepas.org>), are a set of free programs for the analysis of data from microarrays [16]. These softwares are very comprehensive and useful; however, it is necessary to work with human, mouse or other species, because the programs lack an orthologous step beyond the most usual







biological species. Our program offers the possibility to perform orthology searches in other biological species such as fish. The software described here, and the application to one example, show how results from microarray experiments can be integrated into pathways and visualized by using one “gene orthology” step even with uncommon biological species. This enables to drawn pathways in species which are not supported in the KEGG database. The issue here was to derive knowledge of biological relevant patterns in genetic profiling data that were related to the teleost’s immune defences. Accordingly, the role of several genes revealed by the pathway comparison could be defined. Another advantage of the automated procedure used by the Gene2Path software is that no direct supervision is needed and once the process has been initiated, the user can leave the programme running and visualize the results later. In the Agronomic, soils and environmental sciences department have been developed some user friendly software that can be used in industrial companies related with healthy, safety, environment (HSE). Some examples of that software are: Environmental flow diagram (EFD) [17], Soil Heat Calculator Program (SHCP) [18], and Optimize the infiltration parameters in Furrow irrigation using Visual Basic and genetic algorithm [19]. In summary the Gene2Path programme performs an automated search of several databases over 5 steps. (1) The addition of synonyms to identify genes by searching the NCBI database by their IDs. (2) The analysis of the intensity of the gProccesedSignal reading in the one-channel microarray output files (extension “.txt”); (3) An automated search of gene orthology through the homologue, Gene and Unigene NCBI databases, whereby the tool compares nucleotide sequences and comparative 3D models of proteins (constructing an atomic-resolution model of the “target” protein from its amino acid sequence, and producing an experimental 3D structure of a related orthologous protein). (4) The identification of sequences similar to the query set in NCBI nr and EST databases using Xblast. Since other programmes now exist for this step, such as BLAST2GO, we developed an automatic filter to readily visualize the output file (“.xml”) that enables several files to be analyzed at the same time; (5) A search of KEGG database pathway information using orthologous genes (Gene IDs).

## Conclusion

All the software steps were applied to the microarray data we had obtained previously from vaccinated fish: The software proved to be very efficient in terms of automation and data processing. For instance, running a search of different NCBI databases renders 2/3 genes per second (depending on the internet connection speed). The analysis of the data from the vaccinated fish in our example rendered 179 targeted pathways. The Gene2Path software allows the automated searching of NCBI databases and the straightforward visualization of the data retrieved based on a graphic network environment.

## Acknowledgement

We thank Mario García-Lacoba for his advice and helpful comments. This work was funded by Consejo Superior de Investigaciones Científicas (project 2010-20E084) and by Ministerio de Economía y Competitividad, (MINECO) project AGL2010- 18454 of Spain. N. Ballesteros wants to thank the MINECO for a PhD student fellowship.

## References

1. Afshari CA, Hamadeh HK, Bushel PR (2011) The Evolution of Bioinformatics in Toxicology: Advancing Toxicogenomics. *Toxicol Sci* 120: S225-S237.
2. Haab BB (2003) Methods and applications of antibody microarrays in cancer research. *Proteomics* 3: 2116-2122.
3. Salem M, Kenney PB, Rexroad CE, Yao J (2008) Development of a 37 k high-

density oligonucleotide microarray: a new tool for functional genome research in rainbow trout. *J Fish Biol* 72: 2187-2206.

4. Huang D, Pan W (2006) Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics* 22: 1259-1268.
5. Cary MP, Bader GD, Sander C (2005) Pathway information for systems biology. *FEBS Lett* 579: 1815-1820.
6. Pas MFW (2008) A pathway analysis tool for analyzing microarray data for species with low physiological information. *Adv Bioinform* 2008: 1-7.
7. Mortensen SH (1993) The Relevance of Infectious Pancreatic Necrosis Virus (Ipnv) in Farmed Norwegian Turbot (*Scophthalmus-Maximus*). *Aquaculture* 115: 243-252.
8. Rodriguez Saint-Jean S, Vilas Minondo MP, Perez Prieto S (1993) A viral diagnostic survey of Spanish rainbow trout farms: I Sensitivity of four cell lines to wild IPNV isolates. *Bull Eur Assoc Fish Pathol* 13: 119-122.
9. Rodriguez Saint-Jean SP, Borrego JJ, Perez-Prieto SI (2003) Infectious Pancreatic Necrosis Virus: Biology, Pathogenesis, and Diagnostic Methods. *Adv Virus Res Volume* 62: 113-165.
10. Murray AG (2006) Persistence of infectious pancreatic necrosis virus (IPNV) in Scottish salmon (*Salmo salar* L.) farms. *Prev Vet Med* 76: 97-108.
11. Ballesteros NA, Saint-Jean SSR, Encinas PA, Perez-Prieto SI, Coll JM (2012) Oral immunization of rainbow trout to infectious pancreatic necrosis virus (Ipnv) induces different immune gene expression profiles in head kidney and pyloric ceca. *Fish Shellfish Immunol* 33: 174-185.
12. Ballesteros NA (2012) Trout oral VP2 DNA vaccination mimics transcriptional responses occurring after infection with infectious pancreatic necrosis virus (IPNV). *Fish Shellfish Immunol* 33: 1249-1257.
13. Jupiter D, Chen H, VanBuren V (2009) STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data. *BMC Bioinformatics* 10: 332.
14. Haw R, Stein L (2012) Using the reactome database. *Curr. Protoc. Bioinformatics Chapter* 8, Unit8.7.
15. Shannon P (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498-2504.
16. Al-Shahrour F (2008) Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. *Nucleic Acids Res* 36: W341-W346.
17. Valipour M, Morteza Mousavi S, Valipour R, Rezaei E (2012) Air, Water, and Soil Pollution Study in Industrial Units Using Environmental Flow Diagram. *J Basic Appl Sci Res* 2: 12365-12372.
18. Valipour M, Morteza Mousavi S, Valipour R, Rezaei E (2012) SHCP: Soil Heat Calculator Program. *IOSR J Appl Phys* 2: 44-50.
19. Valipour, Mohammad Montazar Asghar A (2012) Optimize of all Effective Infiltration Parameters in Furrow Irrigation Using Visual Basic and Genetic Algorithm Programming. *Aust J Basic Appl Sci* 6: 132.

**Citation:** Ballesteros N, Aguirre N, Coll J, Pérez-Prieto SI, Saint-Jean SR (2015) Gene2Path: A Data Analysis Tool to Study Fish Gene Pathways by Automatic Search of Orthologous Genes. J Aquac Res Development 6: 329. doi:[10.4172/2155-9546.1000329](https://doi.org/10.4172/2155-9546.1000329)