Comparative Analysis of Medaka Genes

Student ID: 47-46911
Name: Wei Qu
Adviser: Pro. Shinichi Morishita

We started the MEDAKA GENOME SEQUENCING PROJECT in 2002. I am involved in this project, entirely responsible for comparative analysis of medaka genes to other fish species' genes and human genes. The assembly of the genome sequence of medaka came over its final stage and 20,141 gene clusters of medaka were identified by a novel method, which took advantage of the comprehensive transcription start site information collected by the high-throughput 5'SAGE method. I performed a comparative analysis on the evolution of medaka genes and provided a further perspective on vertebrate evolution.

Annotation of genes

Among the 20,141 gene clusters predicted, over two thousand genes have no homologues with other fish genes, amphibian genes, human genes, *Takifugu* genome or medaka ESTs. About 67% of Medaka genes possess homologues and 58% have strong corresponding orthologues with human Refseq genes.

Half of the human disease genes have medaka orthologues, which represents the importance of the medaka fish as a model organism in experimental medical science especially in the embryology field.

Experimental evidence for novel genes

Dozens of these novel genes were confirmed by RT-PCR, TA cloning and *in situ* hybridization. Furthermore, a pioneer morpholino-based gene knockdown experiment to elucidate the function of these novel genes was designed and performed.

Evolution of orthologues according to Gene Ontology (GO)

2,292 of the 4,342 medaka-human 1:1 orthologues could be identified with one or multiple GO 'biological process' IDs. Orthologues involved in carbohydrate metabolism, alcohol metabolism, and catabolism were more conserved than those implicated in immune response, transcription, apoptosis, DNA repair and response to stress. This result was similar to that from the comparison of 1:1 chicken-human orthologues that genes related to adaptation to the environment seem to be less conserved in their protein-coding sequences.

Genome size difference in Fish

I computed gene size ratios of reciprocal best matches of medaka and *Takifugu*. The average ratio is about 3 the median ratio is about 2, which is supposed to be an important factor effects

about 1 time increase of genome size from *Takifugu* to medaka. This helped us to understand how the remarkable difference of genome size arose in fishes.

Evolution of Paralogues

I investigated the evolutionary relationships of medaka genes to *Tetraodon* genes and their paralogous pairs, and I found large portion of medaka paralogous pairs are conserved as paralogous pairs *Tetraodon*. These "pair-in-pair" paralogous pairs involved in transport, catabolism, alcohol metabolism, transcription factor NF-*kappa*B, carbohydrate metabolism *etc.* were significantly conserved than those implicated in RNA metabolism, DNA metabolism, transcription, response to DNA damage stimulus, phosphorus metabolism *etc.* , which is quite associated with the fore mentioned analysis of evolution of medaka-human orthologues.

Similarity of expression pattern

I studied the similarity of expression pattern between medaka duplicate genes. A positive correlation of expression levels was found. However, more or less expression divergences were detected in quite a large portion of duplicated genes, which enables tissue or developmental specialization to evolve.

Local gene duplications

Local gene duplications were detected by comparing each gene with its preceding genes and subsequent genes on the chromosome. The number of identified clusters was considerably lower than that in mammals. There are no big clusters of immunoglobulin or olfactory receptor. Keratin genes, which form the scales of fishes, account for a large proportion.

Genome duplication

1,730 pairs of medaka 1:1 duplicated genes were identified, which is a clear signature of whole genome duplication in medaka. With the homologous information of 1:1 human-medaka orthologues, 572 human-medaka synteny blocks were identified. Ancestor chromosomes' blocks on the human chromosomes were detected manually, which is another evidence for the whole genome duplication in fishes.

Transcriptome map

Over one million 5' SAGE (serial analysis of gene expression) tags were collected and 90% of them were mapped to medaka genome successfully. This high-throughput 5'SAGE method provided us both transcription start site information which used to predict genes and genome-wide messenger RNA expression profile as well. I sketched the transcriptional landscape of medaka genome and found domains with highly or weakly expressed genes scattering on the chromosomes. The landscape of expression levels agrees with that of gene density very well.