Database tool

pseudoMap: an innovative and comprehensive resource for identification of siRNA-mediated mechanisms in human transcribed pseudogenes

Wen-Ling Chan¹,², Wen-Kuang Yang³,⁴, Hsien-Da Huang¹,²,* and Jan-Gowth Chang⁵,⁶,⁷,*

¹Institute of Bioinformatics and Systems Biology, ²Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu, ³Cell/Gene Therapy Research Laboratory, Department of Medical Research, China Medical University Hospital, Taichung, ⁴Departments of Biochemistry and Medicine, China Medical University, Taichung, ⁵Center of RNA Biology and Clinical Application, ⁶Department of Laboratory Medicine, China Medical University Hospital, Taichung, and ⁷School of Medicine, China Medical University, Taichung, Taiwan

*Corresponding author: Tel: +886 4 22052121 ext. 2008; Email: d6781@mail.cmuh.org.tw; Fax: +886 4 22031029
Correspondence may also be addressed to Hsien-Da Huang, Tel: +886 3 5712121 ext. 56957; Email: bryan@mail.nctu.edu.tw; Fax: +886 3 5739320

Submitted 31 October 2012; Revised 1 January 2013; Accepted 3 January 2013


RNA interference (RNAi) is a gene silencing process within living cells, which is controlled by the RNA-induced silencing complex with a sequence-specific manner. In flies and mice, the pseudogene transcripts can be processed into short interfering RNAs (siRNAs) that regulate protein-coding genes through the RNAi pathway. Following these findings, we construct an innovative and comprehensive database to elucidate siRNA-mediated mechanism in human transcribed pseudogenes (TPGs). To investigate TPG producing siRNAs that regulate protein-coding genes, we mapped the TPGs to small RNAs (sRNAs) that were supported by publicly deep sequencing data from various sRNA libraries and constructed the TPG-derived siRNA-target interactions. In addition, we also presented that TPGs can act as a target for miRNAs that actually regulate the parental gene. To enable the systematic compilation and updating of these results and additional information, we have developed a database, pseudoMap, capturing various types of information, including sequence data, TPG and cognate annotation, deep sequencing data, RNA-folding structure, gene expression profiles, miRNA annotation and target prediction. As our knowledge, pseudoMap is the first database to demonstrate two mechanisms of human TPGs: encoding siRNAs and decoying miRNAs that target the parental gene. pseudoMap is freely accessible at http://pseudomap.mbc.nctu.edu.tw/.

Database URL: http://pseudomap.mbc.nctu.edu.tw/

Introduction

Pseudogenes are genomic DNA sequences homologous to functional genes yet are not translated into proteins (1). Although pseudogenes are often considered the structurally defective non-functional copies of protein-coding genes, the human genome comprises more numbers of pseudogenes than corresponding functional genes (2). Despite the previous assumption of pseudogenes as genomic fossils, the genome-wide investigations have demonstrated actively transcribed pseudogenes (TPGs) with functional potential (3–12). For instance, TPG of nitric oxide synthase (sNOS) acts as an antisense regulator of neuronal NOS protein synthesis in snails (13, 14). Another study has established that binding of transcriptional repressor to receptor of yapakrin1-p1 could activate the homologous parental gene Mkrn1 (15), despite contradictory...
result was also reported (16). In addition, the TPG of PTENP1 (PTEN), a highly conserved processed pseudogene of tumour suppressor PTEN, acts as a miRNA-decoy by binding to PTEN-targeting miRNAs (17). Moreover, human pseudogene myosin light chain kinase pseudogene 1 is partially duplicated from the original MYLK gene and promotes cancer cell proliferation (18). Above findings clearly suggest that the non-coding RNA products of TPGs may play an important role in biogenesis pathway and functional processes.

The RNA interference (RNAi) is an important component of the RNA modulation pathway and is incorporated into the RNA-induced silencing complex (RISC) with a sequence-specific manner (19). In mice and fruit flies, double-stranded RNAs arising from the antisense/sense transcripts of processed pseudogene, and its cognate gene, or hairpin structures from inversion and duplication, are cut by Dicer into 21 nt endogenous short interfering RNAs (esiRNAs) with the ability to bind RISC and regulate the expression of parental gene (20–25). Such regulatory mechanism in human remains unclear.

To demonstrate that in human, as in animal models, TPGs may generate naturally occurring siRNAs and Piwi interacting RNAs (piRNAs) to regulate the expression of protein-coding genes, we have developed a computational pipeline and constructed a database-pseudoMap, the map for studying pseudogenes. pseudoMap pre-processes the raw data of public microarray and deep sequencing data into gene expression profiles for both TPG and its cognate gene and small RNA (sRNA) profiles for TPG-derived esiRNAs. pseudoMap further combined the gene expression profiles to construct the TPG-derived esiRNA-target interactions (eSTIs). In addition, according to the previous study of pseudogene, PTENP1 exerts a miRNA decoy by binding to cognate-targeting miRNAs (17), and pseudoMap also provided the ‘miRNA regulator’ to elucidate the relationship of TPG and its cognate gene with miRNA target regulation.

**Data generation**

In total, more than 20,000 human pseudogenes and their cognate genes were obtained from the Ensembl Genome Browser (Ensembl 63, GRCh37) (26) using BioMart (http://www.ensembl.org/index.html). Affymetrix GeneChip Human Genome U133A/U133Plus2 is a microarray composed of oligonucleotide probes to measure the level of transcription of each sequence represented, which included transcribed pseudogenes. 1404 pseudogenes have been detectable by this chip, thus considered being transcribed and referred as TPGs. Functional sRNAs (fsRNAs) with sequence length between 18 to 40 nt were collected from the Functional RNA Database (27), which hosts a large collection of known/predicted non-coding RNA sequences from public databases: H-invDB v5.0 (6), FANTOM3 (28), miRBase 17.0 (29, 30), NONCODE v1.0 (31), Rfam v8.1 (32), RNAdb v2.0 (33) and snoRNA-LBME-db rel. 3 (34). The public deep sequencing data from sRNA libraries (35–38) were experimented with on human embryonic stem cells, liver tissues or hepatocellular carcinoma (HCC) tissues. Supplementary Table 1 summarizes the statistics of the deep sequencing data from various sRNA libraries. The genomic sequences were obtained from UCSC hg19 (39). Table 1 lists the integrated databases and tools for mining potential regulators and functions of human TPGs.

**System flow of pseudoMap**

The system flow of pseudoMap is shown in Figure 1, mainly including the collection of datasets such as TPGs, parental
Identification of TPG-derived esiRNAs by public next-generation sequencing data

A computational pipeline was developed to verify the hypotheses that human TPGs may generate esiRNAs to regulate protein-coding genes (Figure 2). An attempt was made to identify the candidates of TPG-derived esiRNAs, by aligning the sequences of TPGs and fsRNAs. These candidates were verified using the deep sequencing data from various sRNA libraries (35–38) experimented with on human embryonic stem cells, liver tissues or HCC tissues. The hairpin structure by Mfold (40) was then determined by using the extended sequences of these candidates of esiRNAs. In pseudoMap, a total of 1232 TPGs may produce esiRNAs, which were profiling by deep sequencing data, within 1404 human TPGs were characterized. The information of these TPGs is shown in Supplementary File 1. The results showed that 4 miRNAs and 326 piRNAs may derive from TPGs. We also found that miRNA has-miR-622 was identified, which was derived from keratin 18 pseudogene 27, located on nt 858–879, as similar as miRBase database. Table 2 summarizes the entire statistical analysis of pseudoMap.

Identification of TPG-derived esiRNA-target interactions

Our previous approach (41) was modified to identify TPG-derived esiRNA targets. Briefly, the esiRNA target
sites within the conserved regions of coding, 5′-UTR and 3′-UTR of genes were identified in 12 metazoan genomes by using three computational approaches, TargetScan (42–44), miRanda (45) and RNAhybrid (46). The minima free energy (MFE) threshold was -20 kcal/mol with a score more than or equal to 150 for miRanda and default parameters for TargetScan and RNAhybrid. The targets were identified using the following criteria: (i) the potential target sites were determined by at least two approaches; (ii) multiple target sites were prioritized and (iii) target sites must be located in accessible regions. Finally, we provided the gene expression profiles of TPG and its cognate gene to construct the eSTIs.

**Gene expression analysis**

The mRNA abundances of TPGs and protein-coding genes were obtained from Gene Expression Omnibus (47), such as GDS596 examined from 79 human physiologically normal tissues (48), GSE2109 examined from 2158 samples with 61 tumour tissues, GSE3526 examined from 353 samples with 65 normal tissues (49) and GSE5364 examined from primary human tumours and adjacent non-tumour tissues, which include 270 tumours and 71 normal-cancer pairs from patients with breast, colon, liver, lung, oesophagal and thyroid cancers (50). Moreover, the Pearson correlation coefficient was computed from TPGs and protein-coding genes.

**Determination of miRNA-target interactions**

According to the study by Poliseno et al. (17), pseudogenes PTENP1 and KRAS1P act as a ‘miRNA decoy’, binding to and thereby reducing the effective cellular concentration of
miRNAs, therefore resulting their cognate genes to escape miRNA-mediated repression. In this study, we analyse the relationships between TPG and its cognate gene with miRNA decoys mechanism to examine miRNA-target interactions (MTIs) by performing a pipeline. First, the parental genes were obtained by mapping the TPGs and genomic sequences with the BLAST (51) program. The MTIs with TPGs and parental genes were then investigated using our previous approach (41). The MFE threshold was $-20 \text{kcal/mol}$ with a score more than or equal to 150 for miRanda and default parameters for TargetScan and RNAhybrid. Finally, the TPGs and their cognates co-regulated by miRNAs were obtained. The miRNA and 3'UTR sequences were obtained from miRBase R18 (29, 30) and Ensembl Genome Browser release 63 (26), respectively. Analysis results indicated that 874 miRNAs with MFE $\leq -20$ and Score $\geq 150$ interact with many possible target sites in 248 TPGs and their cognate genes and might potentially co-regulate this pair of TPG and parental gene (Supplementary File 2).

### Web interface

As a web-based system, pseudoMap can thoroughly identify TPGs, including TPGs act as a miRNA regulators and TPGs-derived eSTIs in humans. There are two ways to access pseudoMap: by browsing the database content or by searching for a particular TPG. Figure 3A displays the

![Figure 3](image-url)

**Figure 3.** Web interface of pseudoMap. (A) Browse interface of pseudoMap illustrates general information of TPGs, miRNA regulators, esiRNAs and gene expression profiles. (B) The miRNA regulator indicates the miRNA decoys mechanisms between TPG and its cognate. (C) Gene expression profiles of TPG and its cognate gene in various experimental conditions. (D) The diagram of esiRNA represents TPG-derived siRNAs as profiled by deep sequencing data. It displays the more fine-grained information of (E) esiRNA-target interaction and (F) RNA folding structure of TPG-derived esiRNA. In addition, pseudoMap also incorporates the external sources, such as (G) UCSC genome browser for a genomic view, GeneCards for gene annotation and miRBase for miRNA annotation.
interface of output results of the browse gateway. The interface contains general information of TPGs, the relationships of TPG and its cognate gene with miRNA-mediated repression termed as ‘miRNA Regulator’, TPG-derived eSTIs named as ‘esiRNA’, and ‘Expression’ showed the gene expression profiles. Figure 3B provides a detailed view of miRNA regulator, which displays more fine-grained information. Above results indicated the relationships between TPGs and cognate genes by a miRNA decoy mechanism such as that observed by Poliseno et al. (17). The ‘Expression’ presents the gene expression profiles of not only distinct TPG and corresponding parental gene but also TPG referenced by cognate in various experimental conditions (Figure 3C). Moreover, the view of esiRNA indicates the TPG-derived esiRNAs and graphical display of deep sequencing data (Figure 3D). The red line represents the TPG and the blue line refers to esiRNAs. We also estimate the eSTIs (Figure 3E) and the RNA folding structure of TPG-derived esiRNA (Figure 3F). All the results and sequences can be downloaded for further experimental tests. In pseudoMap, we also incorporate the external sources, such as UCSC genome browser (39) for a genomic view, GeneCards (52) for gene annotation and miRBase (29) for miRNA annotation (Figure 3G). In addition, pseudoMap also consists of a tutorial and knowledge of pseudogenes.

In the search gateway, the TPG ID, Ensembl ID, TPG symbol and parental gene symbol are allowed for further analysis. Figure 4 displays the interface of output results with the search a particular TPG ID/Ensembl ID/TPG symbol/parental gene symbol. The interface contains the general information of TPG, miRNA regulators, gene expression profiles and TPG-derived esiRNAs.

Figure 4. Search interface of pseudoMap.
Construction and content

In pseudoMap, various databases are integrated and maintained with MySQL (http://www.mysql.com/) relational database management system. While operating on an Apache HTTP server (http://www.apache.org/) and PHP (http://www.php.net/) on a Linux operation system (http://www.linux.com), pseudoMap was constructed using the Smarty template engine (http://www smarty.net). Based on PHP, JavaScript (http://www.javascriptsource.com/), CSS (http://www.w3schools.com/css/) and HTML (http://www.w3schools.com/html/) languages, the web interface enables dynamic MySQL queries with user-friendly graphics. Above software are open source technologies.

Discussion and conclusions

Comparison with other previous databases related to pseudogenes

A few databases have been constructed to explore pseudogenes. In particular, PseudoGene database (53) identifies pseudogenes using various computational methods in genomes; HOPPSIGEN (54) represents the homologous processed pseudogenes shared between the mouse and human genomes that contains location information and potential function; as a web-based system, PseudoGeneQuest (55) identifies novel human pseudogenes based on a user-provided protein sequence; in addition, the University of Iowa’s UI Pseudogenes website contains human pseudogenes and the candidates for gene conversion (56). However, these databases focus on automatic detection of pseudogenes by using a variety of homology-based approaches. Our database, pseudoMap, aims at providing comprehensive resource for genome-wide identifying the functions and regulators of human TPGs. In briefly, there are three major differentiating features from currently public databases of pseudogenes. First, pseudoMap elucidates the relationships of TPG and its cognate gene with miRNA decoys mechanism. Second, to explore the interaction of TPG and its parental gene, pseudoMap provides the gene expression profiles of TPG and its cognate gene in various experimental conditions. Third, pseudoMap curates the TPG-derived esiRNAs, which supported by deep sequencing data, as well as their interacting gene targets in the human genome. Table 3 lists the detailed comparisons of pseudoMap with other previous databases related to pseudogenes.

Applications

PseudoMap provides two major applications. One is the non-coding RNA products of TPGs, as like animal models, that may generate esiRNAs to regulate protein-coding genes in humans. In this process, pseudoMap supplies next-generation sequencing data from sRNA libraries to
support the candidates of TPG-derived esiRNAs and gene expression profiles to verify the target interactions, respectively. Another application is that both the gene and pseudogene contain miRNA target sites, if the pseudogene competes for the freely available repressor molecules that would be free the gene to reduce the miRNA-mediated repression. Another words, the pseudogene may act as a ‘miRNA decoy’ to release the repression of its cognate gene. pseudoMap provides another insight into the pathway of MTIs with TPG-mediated mechanism.

Conclusion
In this study, we performed a computational pipeline to identify TPG-derived esiRNAs-target interactions and constructed a comprehensive database to represent the potential functions and regulators of TPGs in human. To our knowledge, the pseudoMap is the first database to identify TPGs to enable biologists and bioinformaticians to elucidate two major discoveries, the relationships between TPG and its cognate gene with miRNA decoyed mechanisms and TPG-derived eSTIs. Efforts are underway in our laboratory to expand the methods used in pseudoMap to other species such as mice, fruit flies and plants. The pseudoMap will be updated frequently by continuingly surveying experimentally validated sRNAs and will be maintained with a long-term support from National Chiao Tung University and National Science Council at Taiwan. This novel and creative resource is now freely available at http://pseudomap.mbc.nctu.edu.tw/.

Supplementary data
Supplementary Data are available at Database online.

Funding

Conflict of interest statement. None declared.

References


