Database tool

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

Wen-Ling Chan[1,2], Wen-Kuang Yang[3,4], Hsien-Da Huang[1,2,*] and Jan-Gowth Chang[5,6,7,*]

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

Citation details: Chan,W.-L., Yang,W.-K., Huang,H.-D. *et al*. pseudoMap: an innovative and comprehensive resource for identification of siRNA-mediated mechanisms in human transcribed pseudogenes. *Database* (2013) Vol. 2013: article ID bat001; doi:10.1093/database/bat001.

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 instant, TPG of nitric oxide synthase (*ψNOS*) 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 *ψmakorin1-p1* could activate the homologous parental gene *Mkrn1* (15), despite contradictory result was also reported (16). In addition, the TPG of *PTENP1* (ψ*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 embryo 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

Table 1. Supported databases and tools in pseudoMap

Integrated database or tools	Dataset	Description
miRBase (29, 30)	miRNA annotation	This database not only provides published miRNA sequences and annotations but also supplies known/predict targets
Functional RNA Database (27)	sRNA annotation	A database to support mining and annotation of functional RNAs
Ensembl Genome Browser (26)	Pseudogene, protein-coding gene	It produces genome databases for vertebrates and other eukaryotic species
UCSC Genome Browser (39)	Conserved region and Genomic view of genes	This browser provides a rapid and reliable display of any requested portion of genomes at any scale, together with dozens of aligned annotation tracks
GeneCards (52)	Gene annotation	GeneCards is a searchable, integrated, database of human genes that provides concise genomic-related information of all known and predicted human genes
Mfold (40)	RNA folding tool	Folding RNA structure
GEO (47)	Gene expression profiles and deep sequencing data	A public functional genomics data
BLAST (51)	Sequence alignment tool	BLAST finds regions of similarity between biological sequences

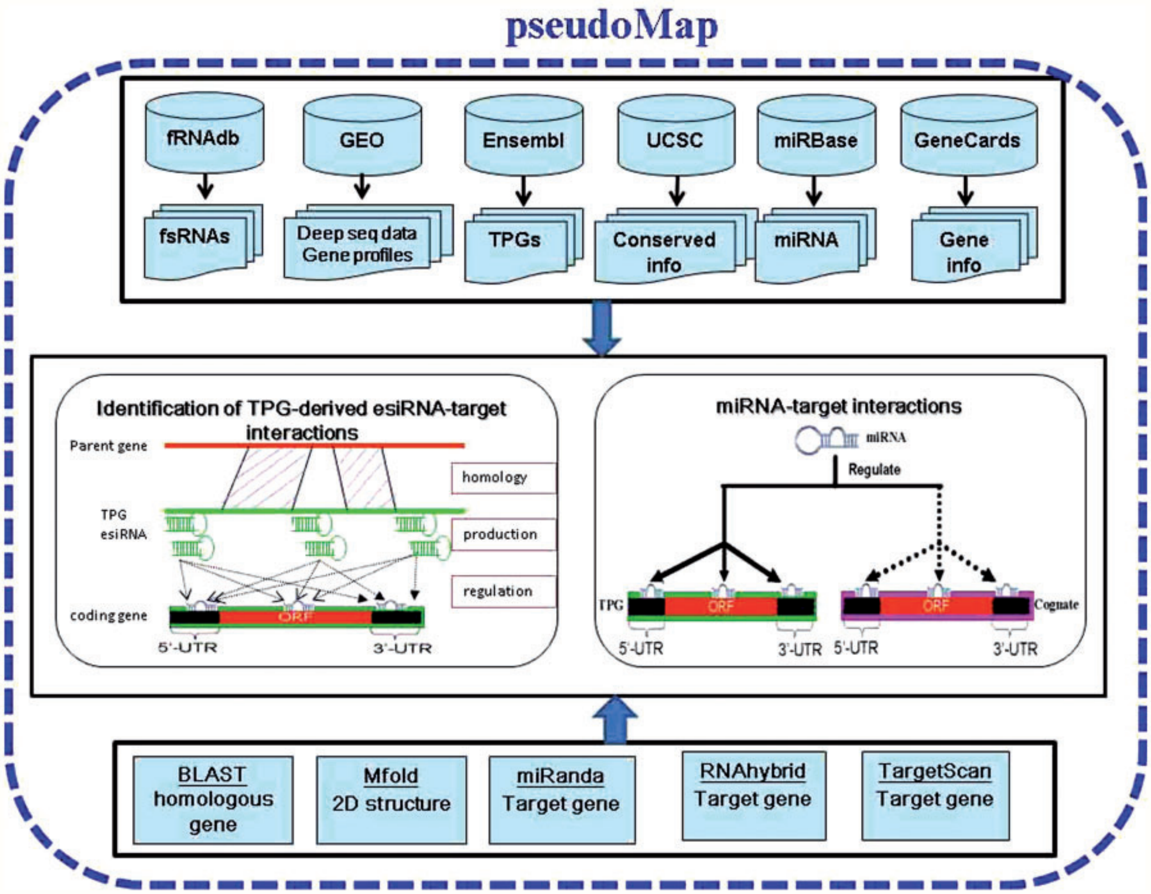


Figure 1. System flow of pseudoMap. **The system flow of pseudoMap mainly includes** the collection of datasets such as TPGs, parental genes, miRNAs, piRNAs, sRNA deep sequencing data and expression profiles; integration of various tools and identification of functions and regulations of TPGs. Based on a genome-wide computational pipeline of sequence-alignment approaches and gene expression profiles, this work constructed pseudoMap database for elucidation of two major discoveries: TPG-derived esiRNA-target interaction and miRNA-decoy mechanism of TPGs.

genes, fsRNAs, sRNA deep sequencing data, expression profiles, integration of various tools and identification of functions and regulations of TPGs. Based on a genome-wide computational pipeline of sequence-alignment approaches, this work constructed pseudoMap database for elucidation of two major discoveries: TPG-derived eSTI and miRNA-decoy mechanism of TPGs. The detailed analyses are described below.

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 embryo 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

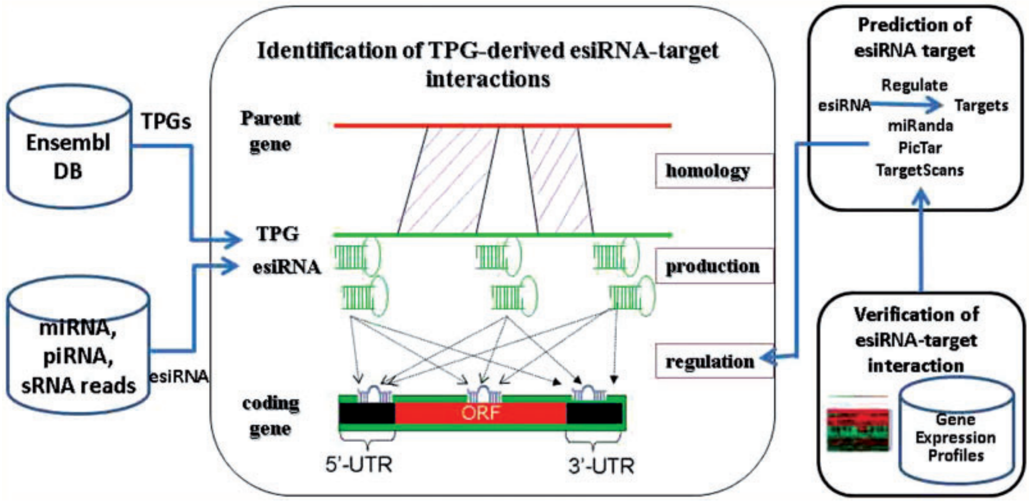


Figure 2. Computational pipeline for identification of TPG-derived esiRNA-target interactions.

Table 2. Summarizes the entire statistical analysis of pseudoMap

Dataset	Counts
No. of miRNA regulators	5771/1014[a]
No. of TPG-derived miRNAs	4
No. of TPG-derived piRNAs	326
Deep sequencing data for profiling TPG-derived esiRNAs	
Human embryo stem cell—hB	247
Human embryo stem cell—hESC	553
Human embryo stem cell—hues6	190
Human embryo stem cell—hues6NP	81
Human embryo stem cell—hues6Neuron	16
HBV(+) adjacent tissue sample 1	917
HBV(+) adjacent tissue sample 2	4377
HBV(+) distal tissue sample 1	1011
HBV(+) HCC tissue sample 1	1281
HBV(+) HCC tissue sample 2	2649
HBV-infected liver tissue	3056
HBV(+) side tissue sample 1	1087
HCV(+) adjacent tissue sample	14 297
HCV(+) HCC tissue sample	9277
HBV(−) HCV(−) adjacent tissue sample	2324
HBV(−) HCV(−) HCC tissue sample	6579
Human normal liver tissue sample 1	1220
Human normal liver tissue sample 2	1290
Human normal liver tissue sample 3	1209
Severe chronic hepatitis B liver tissue	1247

[a]1014 distinct miRNAs involved in 5771 miRNA regulators.

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

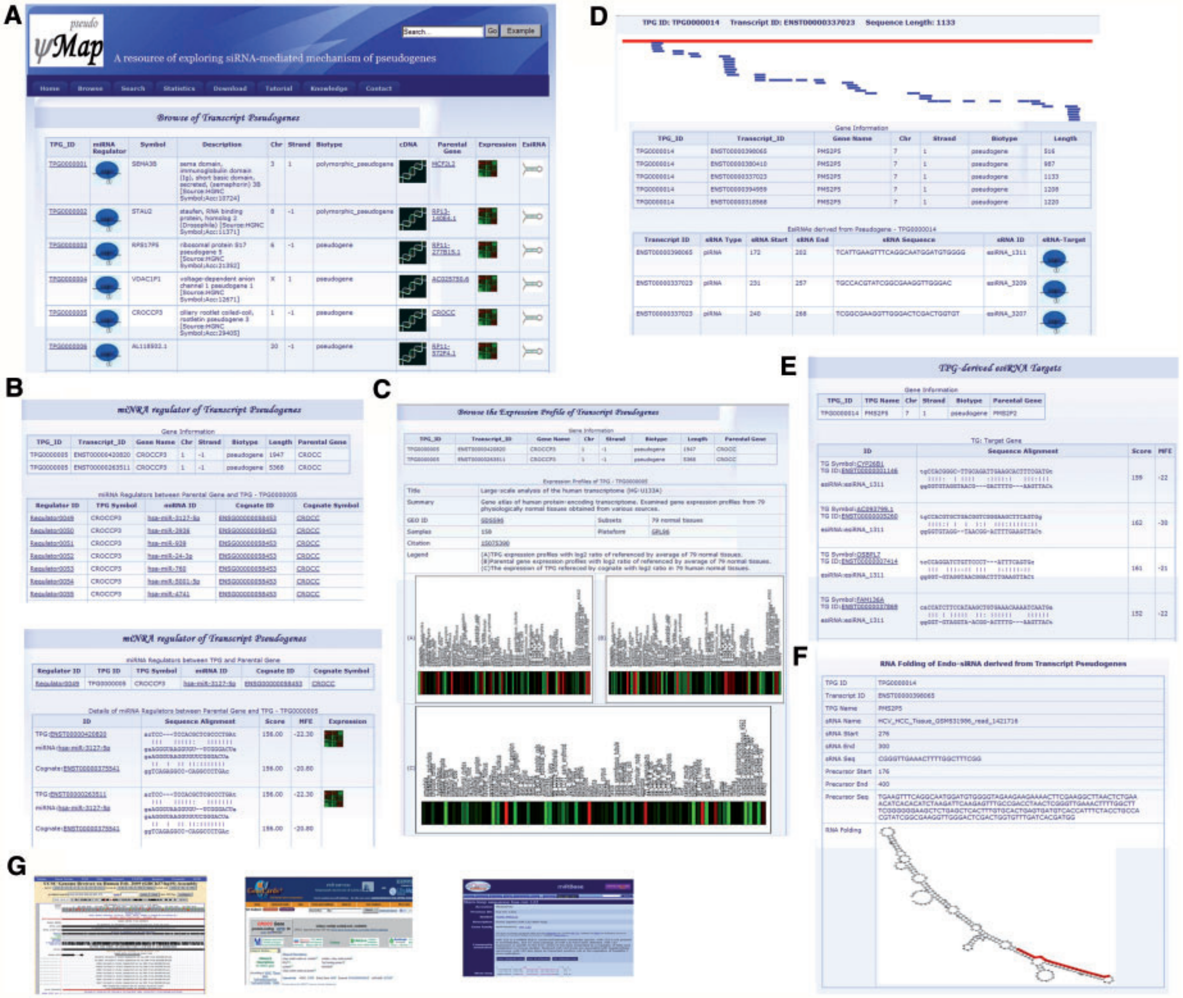
The mRNA abundances of TPGs and protein-coding genes were obtained from Gene Expression Omnibus (47), such as GDS596 examined from 79 human physiologically normal tissues (48), GSE2109 examined from 2158 samples with 61 tumour tissues, GSE3526 examined from 353 samples with 65 normal tissues (49) and 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 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 ≤ −20 and Score ≥ 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



Browse of Transcript Pseudogenes

TPG_ID	miRNA Regulator	Symbol	Description	Chr	Strand	Biotype	cDNA	Parental Gene	Expression	EsiRNA
TPG0000001		SEMA3B	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3B [Source:HGNC Symbol;Acc:10724]	3	1	polymorphic_pseudogene		MCF2L2		
TPG0000002		STAU2	staufen, RNA binding protein, homolog 2 (Drosophila) [Source:HGNC Symbol;Acc:11371]	8	-1	polymorphic_pseudogene		RP13-140E4.1		
TPG0000003		RPS17P5	ribosomal protein S17 pseudogene 5 [Source:HGNC Symbol;Acc:21392]	6	-1	pseudogene		RP11-277B15.1		
TPG0000004		VDAC1P1	voltage-dependent anion channel 1 pseudogene 1 [Source:HGNC Symbol;Acc:12671]	X	1	pseudogene		AC025750.6		
TPG0000005		CROCCP3	ciliary rootlet coiled-coil, rootletin pseudogene 3 [Source:HGNC Symbol;Acc:29405]	1	-1	pseudogene		CROCC		
TPG0000006		AL118502.1		20	-1	pseudogene		RP11-572P4.1		

miRNA regulator of Transcript Pseudogenes

Gene Information

TPG_ID	Transcript_ID	Gene Name	Chr	Strand	Biotype	Length	Parental Gene
TPG0000005	ENST00000420820	CROCCP3	1	-1	pseudogene	1947	CROCC
TPG0000005	ENST00000263511	CROCCP3	1	-1	pseudogene	5368	CROCC

miRNA Regulators between Parental Gene and TPG - TPG0000005

Regulator ID	TPG Symbol	miRNA ID	Cognate ID	Cognate Symbol
Regulator0049	CROCCP3	hsa-miR-3127-5p	ENSG00000058453	CROCC
Regulator0050	CROCCP3	hsa-miR-3936	ENSG00000058453	CROCC
Regulator0051	CROCCP3	hsa-miR-938	ENSG00000058453	CROCC
Regulator0052	CROCCP3	hsa-miR-24-3p	ENSG00000058453	CROCC
Regulator0053	CROCCP3	hsa-miR-760	ENSG00000058453	CROCC
Regulator0054	CROCCP3	hsa-miR-5001-5p	ENSG00000058453	CROCC
Regulator0055	CROCCP3	hsa-miR-4741	ENSG00000058453	CROCC

miRNA regulator of Transcript Pseudogenes

miRNA Regulators between TPG and Parental Gene

Regulator ID	TPG ID	TPG Symbol	miRNA ID	Cognate ID	Cognate Symbol
Regulator0049	TPG0000005	CROCCP3	hsa-miR-3127-5p	ENSG00000058453	CROCC

Details of miRNA Regulators between Parental Gene and TPG - TPG0000005

ID	Sequence Alignment	Score	MFE	Expression
TPG:ENST00000420820 miRNA:hsa-miR-3127-5p	[illegible]	156.00	-22.30	
Cognate:ENST00000375541	[illegible]	158.00	-20.80	
TPG:ENST00000263511 miRNA:hsa-miR-3127-5p	[illegible]	156.00	-22.30	
Cognate:ENST00000375541	[illegible]	158.00	-20.80	

Browse the Expression Profile of Transcript Pseudogenes

Gene Information

TPG_ID	Transcript_ID	Gene Name	Chr	Strand	Biotype	Length	Parental Gene
TPG0000005	ENST00000420820	CROCCP3	1	-1	pseudogene	1947	CROCC
TPG0000005	ENST00000263511	CROCCP3	1	-1	pseudogene	5368	CROCC

Expression Profiles of TPG - TPG0000005

Title	Large-scale analysis of the human transcriptome (HG-U133A)		
Summary	Gene atlas of human protein-encoding transcriptome. Examined gene expression profiles from 79 physiologically normal tissues obtained from various sources.		
GEO ID	GDS596	Subsets	79 normal tissues
Samples	158	Plateform	GPL96
Citation	15075390		
Legend	(A)TPG expression profiles with log2 ratio of referenced by average of 79 normal tissues. (B)Parental gene expression profiles with log2 ratio of referenced by average of 79 normal tissues. (C)The expression of TPG refereced by cognate with log2 ratio in 79 human normal tissues.		

TPG ID: TPG0000014 Transcript ID: ENST00000337023 Sequence Length: 1133

Gene Information

TPG_ID	Transcript_ID	Gene Name	Chr	Strand	Biotype	Length
TPG0000014	ENST00000398065	PMS2P5	7	1	pseudogene	516
TPG0000014	ENST00000380410	PMS2P5	7	1	pseudogene	987
TPG0000014	ENST00000337023	PMS2P5	7	1	pseudogene	1133
TPG0000014	ENST00000394989	PMS2P5	7	1	pseudogene	1208
TPG0000014	ENST00000318568	PMS2P5	7	1	pseudogene	1220

EsiRNAs derived from Pseudogene - TPG0000014

Transcript ID	sRNA Type	sRNA Start	sRNA End	sRNA Sequence	sRNA ID	sRNA-Target
ENST00000398065	piRNA	172	202	TCATTGAAGTTTCAGGCAATGGATGTGGGG	esiRNA_1311	
ENST00000337023	piRNA	231	257	TGCCACGTATCGGCGAAGGTTGGGAC	esiRNA_3209	
ENST00000337023	piRNA	240	268	TCGGCGAAGGTTGGGACTCGACTGGTGT	esiRNA_3207	

TPG-derived esiRNA Targets

Gene Information

TPG_ID	TPG Name	Chr	Strand	Biotype	Parental Gene
TPG0000014	PMS2P5	7	1	pseudogene	PMS2P2

TG: Target Gene

ID	Sequence Alignment	Score	MFE
TG Symbol:CYP26B1 TG ID:ENST00000001146 esiRNA:esiRNA_1311	[illegible]	159	-22
TG Symbol:AC093799.1 TG ID:ENST00000005260 esiRNA:esiRNA_1311	[illegible]	162	-30
TG Symbol:OSBPL7 TG ID:ENST00000007414 esiRNA:esiRNA_1311	[illegible]	161	-21
TG Symbol:FAM126A TG ID:[illegible] esiRNA:esiRNA_1311	[illegible]	152	-22

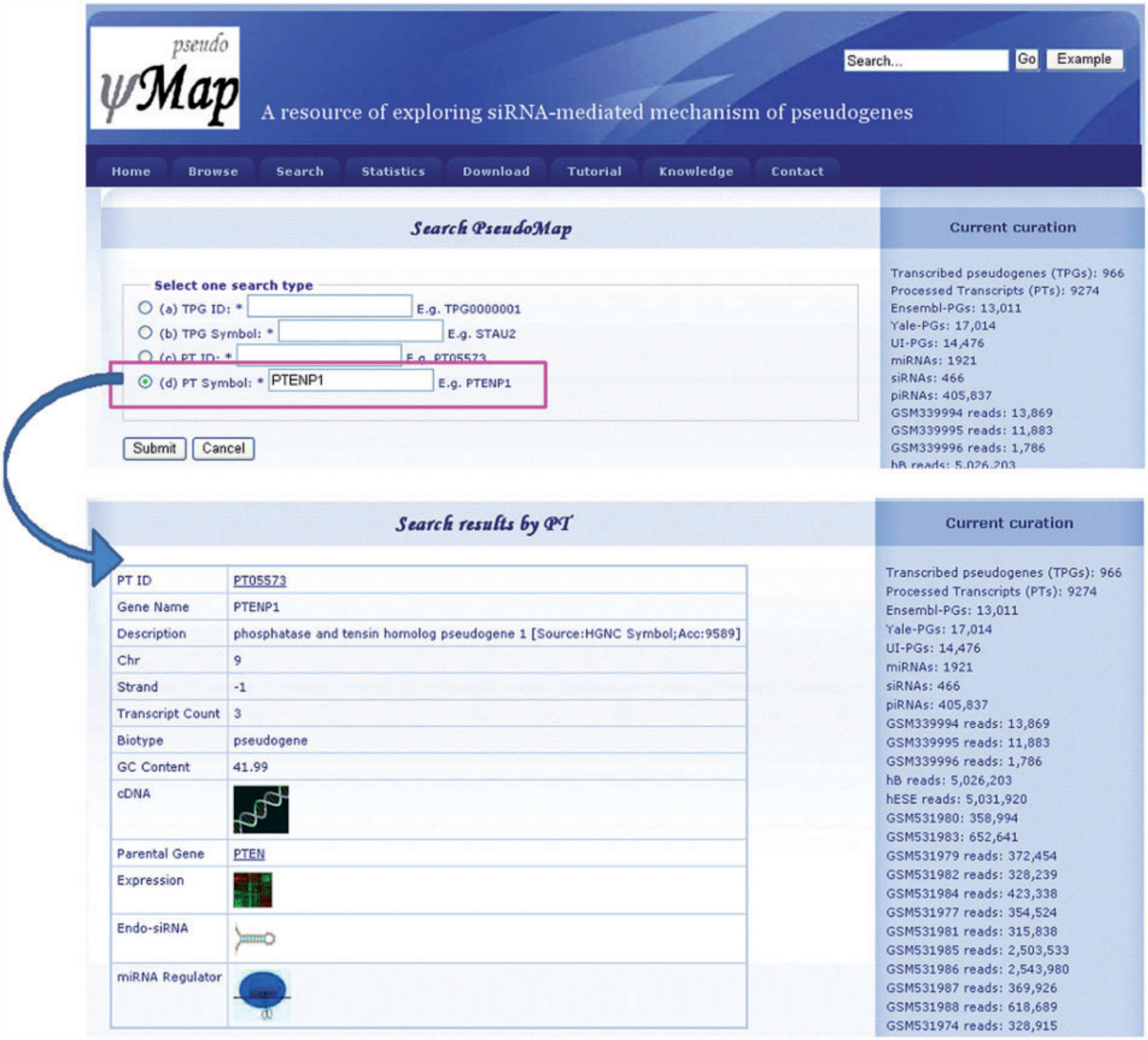
RNA Folding of Endo-siRNA derived from Transcript Pseudogenes

TPG ID	TPG0000014
Transcript ID	ENST00000398065
TPG Name	PMS2P5
sRNA Name	HCV_HCC_Tissue_GSM531986_read_1421718
sRNA Start	276
sRNA End	300
sRNA Seq	CGGGTTGAAACTTTTGGCTTTCGG
Precursor Start	176
Precursor End	400
Precursor Seq	TGAAGTTTCAGGCAATGGATGTGGGGTAGAAGAAGAAAACTTCGAAGGCTTAACTCTGAA ACATCACACATCTAAGATTCAAGAGTTTGCCGACCTAACTCGGGTTGAAACTTTTGGCTT TCGGGGGGAAGCTCTGAGCTCACTTTGTGCACTGAGTGATGTCACCATTTCTACCTGCCA CGTATCGGCGAAGGTTGGGACTCGACTGGTGTTTGATCACGATGG
RNA Folding	

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.



Search PseudoMap

Select one search type

○ (a) TPG ID: * E.g. TPG0000001
○ (b) TPG Symbol: * E.g. STAU2
○ (c) PT ID: * E.g. PT05573
● (d) PT Symbol: * PTENP1 E.g. PTENP1

Current curation

Transcribed pseudogenes (TPGs): 966
Processed Transcripts (PTs): 9274
Ensembl-PGs: 13,011
Yale-PGs: 17,014
UI-PGs: 14,476
miRNAs: 1921
siRNAs: 466
piRNAs: 405,837
GSM339994 reads: 13,869
GSM339995 reads: 11,883
GSM339996 reads: 1,786
hB reads: 5,026,203

Search results by PT

PT ID	PT05573
Gene Name	PTENP1
Description	phosphatase and tensin homolog pseudogene 1 [Source:HGNC Symbol;Acc:9589]
Chr	9
Strand	-1
Transcript Count	3
Biotype	pseudogene
GC Content	41.99
cDNA	
Parental Gene	PTEN
Expression	
Endo-siRNA	
miRNA Regulator	

Current curation

Transcribed pseudogenes (TPGs): 966
Processed Transcripts (PTs): 9274
Ensembl-PGs: 13,011
Yale-PGs: 17,014
UI-PGs: 14,476
miRNAs: 1921
siRNAs: 466
piRNAs: 405,837
GSM339994 reads: 13,869
GSM339995 reads: 11,883
GSM339996 reads: 1,786
hB reads: 5,026,203
hESE reads: 5,031,920
GSM531980: 358,994
GSM531983: 652,641
GSM531979 reads: 372,454
GSM531982 reads: 328,239
GSM531984 reads: 423,338
GSM531977 reads: 354,524
GSM531981 reads: 315,838
GSM531985 reads: 2,503,533
GSM531986 reads: 2,543,980
GSM531987 reads: 369,926
GSM531988 reads: 618,689
GSM531974 reads: 328,915

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

Table 3. Comparisons of pseudoMap with currently public databases of pseudogenes

Supported features	pseudoMap (our database)	PseudoGene database	UI pseudogene	Hoppsigen
Web interface	http://pseudomap.mbc.nctu.edu.tw/	http://www.pseudogene.org/	https://genome.uiowa.edu/pseudogenes/	http://pbil.univ-lyon1.fr/databases/hoppsigen.html
Description	pseudoMap provides a comprehensive resource for genome-wide identifying the functions and regulators of human pseudogenes.	This site contains a comprehensive database of identified pseudogenes, utilities used to find pseudogenes, various publication data sets and a pseudogene knowledgebase.	This site serves as a repository for all pseudogenes in the human genome and provides a ranked list of human pseudogenes that have been identified as candidates for gene conversion.	Hoppsigen is a nucleic database of homologous processed pseudogenes.
Species supported	Human	Eukaryote and prokaryote	Human	Human
Sequence download	Yes	Yes	Yes	Yes
Pseudogene information	Yes	Yes	Yes	Yes
Parental gene information	Yes	Yes	—	—
Knowledge of pseudogenes	Yes	Yes	—	Yes
miRNA–pseudogene interactions	Yes	—	—	—
miRNA–parental gene interactions	Yes	—	—	—
Gene expression profiles	Yes (both pseudogene and its parental gene)	—	—	—
Pseudogene-derived siRNAs	Yes	—	—	—
Deep sequencing data for profiling TPG-derived siRNAs	Yes	—	—	—

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

National Science Council of the Republic of China [NSC 99-2320-B-037-006-MY3 to J.G.C., NSC 98-2314-B039-010MY3 to W.K.Y., NSC 98-2311-B-009-004-MY3 to H.D.H., NSC 99-2627-B-009-003 to H.D.H., NSC 101-2311-B-009-003-MY3 to H.D.H. and NSC 100-2627-B-009-002 to H.D.H.]; UST-UCSD International Center of Excellence in Advanced Bio-engineering sponsored by the Taiwan National Science Council I-RiCE Program [NSC 101-2911-I-009 -101 - to H.D.H., in part]; Veterans General Hospitals and University System of Taiwan (VGHUST) Joint Research Program [VGHUST101-G5-1-1 to H.D.H., in part**];** MOE ATU [in part**]**. Funding for open access charge: National Science Council of the Republic of China**.**

Conflict of interest statement. None declared.

References

1. Mighell,A.J., Smith,N.R., Robinson,P.A. *et al*. (2000) Vertebrate pseudogenes. *FEBS Lett.*, **468**, 109–114.
2. Torrents,D., Suyama,M., Zdobnov,E. *et al*. (2003) A genome-wide survey of human pseudogenes. *Genome Res.*, **13**, 2559–2567.
3. Balasubramanian,S., Zheng,D., Liu,Y.J. *et al*. (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol.*, **10**, R2.
4. Harrison,P. and Yu,Z. (2007) Frame disruptions in human mRNA transcripts, and their relationship with splicing and protein structures. *BMC Genomics*, **8**, 371.
5. Harrison,P.M., Zheng,D., Zhang,Z. *et al*. (2005) Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res.*, **33**, 2374–2383.
6. Imanishi,T., Itoh,T., Suzuki,Y. *et al*. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
7. Khachane,A.N. and Harrison,P.M. (2009) Assessing the genomic evidence for conserved transcribed pseudogenes under selection. *BMC Genomics*, **10**, 435.
8. Vinckenbosch,N., Dupanloup,I. and Kaessmann,H. (2006) Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl Acad. Sci. USA*, **103**, 3220–3225.
9. Zheng,D., Zhang,Z., Harrison,P.M. *et al*. (2005) Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J. Mol. Biol.*, **349**, 27–45.
10. Zheng,D., Frankish,A., Baertsch,R. *et al*. (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res.*, **17**, 839–851.
11. McCarrey,J.R. and Riggs,A.D. (1986) Determinator-inhibitor pairs as a mechanism for threshold setting in development: a possible function for pseudogenes. *Proc. Natl Acad. Sci. USA*, **83**, 679–683.
12. Zou,C., Lehti-Shiu,M.D., Thibaud-Nissen,F. *et al*. (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.*, **151**, 3–15.
13. Korneev,S. and O'Shea,M. (2002) Evolution of nitric oxide synthase regulatory genes by DNA inversion. *Mol. Biol. Evol.*, **19**, 1228–1233.
14. Korneev,S.A., Park,J.H. and O'Shea,M. (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.*, **19**, 7711–7720.
15. Hirotsune,S., Yoshida,N., Chen,A. *et al*. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, **423**, 91–96.
16. Gray,T.A., Wilson,A., Fortin,P.J. *et al*. (2006) The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. *Proc. Natl Acad. Sci. USA*, **103**, 12039–12044.
17. Poliseno,L., Salmena,L., Zhang,J. *et al*. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, **465**, 1033–1038.
18. Han,Y.J., Ma,S.F., Yourek,G. *et al*. (2011) A transcribed pseudogene of MYLK promotes cell proliferation. *FASEB J.*, **25**, 2305–2312.
19. Kim,D.H. and Rossi,J.J. (2007) Strategies for silencing human disease using RNA interference. *Nat. Rev. Genet.*, **8**, 173–184.
20. Czech,B., Malone,C.D., Zhou,R. *et al*. (2008) An endogenous small interfering RNA pathway in *Drosophila*. *Nature*, **453**, 798–802.

21. Ghildiyal,M., Seitz,H., Horwich,M.D. *et al*. (2008) Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science*, **320**, 1077–1081.
22. Kawamura,Y., Saito,K., Kin,T. *et al*. (2008) *Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, **453**, 793–797.
23. Okamura,K., Chung,W.J., Ruby,J.G. *et al*. (2008) The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, **453**, 803–806.
24. Tam,O.H., Aravin,A.A., Stein,P. *et al*. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.
25. Watanabe,T., Totoki,Y., Toyoda,A. *et al*. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, **453**, 539–543.
26. Flicek,P., Amode,M.R., Barrell,D. *et al*. (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
27. Mituyama,T., Yamada,K., Hattori,E. *et al*. (2009) The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.*, **37**, D89–D92.
28. Carninci,P., Kasukawa,T., Katayama,S. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
29. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, **39**, D152–D157.
30. Griffiths-Jones,S., Saini,H.K., van Dongen,S. *et al*. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
31. He,S., Liu,C., Skogerbo,G. *et al*. (2008) NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.*, **36**, D170–D172.
32. Griffiths-Jones,S., Moxon,S., Marshall,M. *et al*. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
33. Pang,K.C., Stephen,S., Dinger,M.E. *et al*. (2007) RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.*, **35**, D178–D182.
34. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
35. Morin,R.D., O'Connor,M.D., Griffith,M. *et al*. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.
36. Seila,A.C., Calabrese,J.M., Levine,S.S. *et al*. (2008) Divergent transcription from active promoters. *Science*, **322**, 1849–1851.
37. Yeo,G.W., Xu,X., Liang,T.Y. *et al*. (2007) Alternative splicing events identified in human embryonic stem cells and neural progenitors. *PLoS Comput. Biol.*, **3**, 1951–1967.
38. Hou,J., Lin,L., Zhou,W. *et al*. (2011) Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma. *Cancer Cell*, **19**, 232–243.
39. Kent,W.J., Sugnet,C.W., Furey,T.S. *et al*. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
40. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
41. Hsu,S.D., Chu,C.H., Tsou,A.P. *et al*. (2008) miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Res.*, **36**, D165–D169.
42. Friedman,R.C., Farh,K.K., Burge,C.B. *et al*. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
43. Grimson,A., Farh,K.K., Johnston,W.K. *et al*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
44. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
45. John,B., Enright,A.J., Aravin,A. *et al*. (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.
46. Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.
47. Barrett,T. and Edgar,R. (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol.*, **411**, 352–369.
48. Su,A.I., Wiltshire,T., Batalov,S. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
49. Roth,R.B., Hevezi,P., Lee,J. *et al*. (2006) Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, **7**, 67–80.
50. Yu,K., Ganesan,K., Tan,L.K. *et al*. (2008) A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genet.*, **4**, e1000129.
51. Altschul,S.F., Gish,W., Miller,W. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
52. Stelzer,G., Dalah,I., Stein,T.I. *et al*. (2011) In-silico human genomics with GeneCards. *Hum. Genomics*, **5**, 709–717.
53. Karro,J.E., Yan,Y., Zheng,D. *et al*. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.*, **35**, D55–D60.
54. Khelifi,A., Duret,L. and Mouchiroud,D. (2005) HOPPSIGEN: a database of human and mouse processed pseudogenes. *Nucleic Acids Res.*, **33**, D59–D66.
55. Ortutay,C. and Vihinen,M. (2008) PseudoGeneQuest—service for identification of different pseudogene types in the human genome. *BMC Bioinformatics*, **9**, 299.
56. Bischof,J.M., Chiang,A.P., Scheetz,T.E. *et al*. (2006) Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Hum. Mutat.*, **27**, 545–552.