Article Chinese Science Bulletin

Marine Biology

March 2013 Vol.58 No.9: 1010–1018
doi: 10.1007/s11434-012-5174-3

Identification and bioinformatics analysis of pseudogenes from whole genome sequence of *Phaeodactylum tricornutum*

JI ChangMian[1†], HUANG AiYou[2,3†], LIU WenLing[1], PAN GuangHua[1] & WANG GuangCe[2*]

[1] *Tianjin Key Laboratory of Marine Resources and Chemistry, College of Marine Science and Engineering, Tianjin University of Science and Technology, Tianjin 300457, China;*

[2] *Institute of Oceanology, Chinese Academy of Sciences (IOCAS), Qingdao 266071, China;*

[3] *Graduate University of Chinese Academy of Sciences, Beijing 100049, China*

Received October 10, 2011; accepted March 20, 2012; published online November 28, 2012

Pseudogenes share sequence similarities with functional genes, but in general they have lost their protein-coding ability. The identification of pseudogenes is a very important step in genome annotation. *Phaeodactylum tricornutum* is a marine diatom that is rich in polyunsaturated fatty acids (PUFAs). The genome of *P. tricornutum* has been completely sequenced. To identify pseudogenes in *P. tricornutum*, we developed a pipeline to discover and characterize pseudogenes. We identified a total of 1654 'true' processed pseudogenes, 714 duplicated pseudogenes and 4729 pseudogene fragments. The results of the bioinformatics analysis indicated that the genome sequence of *P. tricornutum* contained many pseudogenes and pseudogene fragments.

***Phaeodactylum tricornutum*, pseudogene, pseudogene fragment, bioinformatics**

Citation: Ji C M, Huang A Y, Liu W L, et al. Identification and bioinformatics analysis of pseudogenes from whole genome sequence of *Phaeodactylum tricornutum*. Chin Sci Bull, 2013, 58: 1010–1018, doi: 10.1007/s11434-012-5174-3

Pseudogenes are sequences that share close similarities to one or more paralogous functional genes, but in general are not expressed in the cell [1,2]. Based on their origin, pseudogenes have been divided into two categories: duplicated pseudogenes and processed pseudogenes. The former are produced by genomic DNA duplication or by the unequal exchange of chromosomes and retain the basic structure of functional genes, for example, promoters and introns [3–5]. The latter originate from the reverse transcription of mRNA that was inserted back into the genomes and processed pseudogenes rarely contain promoters or introns [1,6]; some of them, however, do retain a part of the poly A tail which may not have been fully degraded. The identification of pseudogenes in genomes is important for the accurate identification and annotation of the functional genes, for the evolutionary analysis of genomes and functional genes [7], and for the determination of the function of the pseudogenes [8–10]. Pseudogenes are commonly found in the genomes of a great variety of species, having been found in *E. coli* [11], yeast [12], *Arabidopsis* [13,14], rice [15], nematodes [16], *Drosophila* [17], mice [18], and humans [19,20]. It has been predicted that the human genome encodes from 22000 to 75000 genes, about 22% of which are pseudogenes [21,22].

Plankton consists mainly of diatoms and planktonic diatoms are widely distributed in oceans, lakes, rivers and other water areas where they play a very important role in the carbon fixation and inorganic circulation cycles, particularly the silicon cycle. Nearly 20% of primary productivity is contributed by diatoms [23–26]. Diatoms have a special evolutionary position in that they originated from secondary endosymbiosis. These algae, therefore, make good material for evolutionary studies [27–31]. *Phaeodactylum tricornutum* is a pennate diatom that is used in the aquaculture industry as bait for shellfish and shrimp because of the abundant unsaturated fatty acids that it produces. The 33 *P. tricornutum* chromosomes and the chloroplast genome have been fully sequenced [32]. Because the genome was found

† These authors contributed equally to this work.

* Corresponding author (email: gcwang@qdio.ac.cn)

to contain many functional genes derived from animals, plants and bacteria, the *P. tricornutum* genome has been reported to be more like an animal genome than a plant genome [32]. The *P. tricornutum* genome size is 27 Mb encoding 10402 genes [32]; approximately 400 functional genes/Mb. In contrast, there are only 7–25 functional genes/Mb in the human genome [19,20], much lower than in the *P. tricornutum* genome. The high functional gene density in the *P. tricornutum* genome could mean that there are only a few or no pseudogenes in the *P. tricornutum* genome.

In this study, we used bioinformatics methods based on C# in the visual studio.net platform and the SQL Server platform to create a procedure to comprehensively search for pseudogenes from the *P. tricornutum* genome [19]. We investigated the correlations between the chromosome length, the GC content, the completeness, and the identity of pseudogenes and the corresponding functional genes. The results showed that there were many processed pseudogenes and pseudogene fragments in the *P. tricornutum* genome. These data will make a significant contribution to in-depth analyses of the *P. tricornutum* genome and its functional genes.

1 Materials and methods

1.1 Source data sets

The publicly available data that were used in this study were the genomic sequence, a functional protein data set and the cDNA data set that corresponds to the functional protein dataset. The genomic data were downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/genome/) and included the sequences of the 33 *P. tricornutum* chromosomes (Assembly name: ASM15095v1). The functional protein data set (Phatr2_chromosomes_geneModels_FilteredModels2_aa.fa sta.gz) and the corresponding cDNA data set (Phatr2_ chromosomes_geneModels_FilteredModels2_nt.fasta.gz) were downloaded from the U.S. Department of Energy Joint Genomes Institute (DOE-JGI) (http://genome.jgipsf.org/Phatr2/ Phatr2.download.ftp.html). The filtered models (“best”) that correspond to the non-redundant protein and cDNA data sets (FilteredModels2 data) were used in this study.

1.2 Bioinformatics methods and statistical analysis

All the bioinformatics methods used in this paper were developed using C# on the visual studio.net platform and SQL on the SQL Server platform. Statistical analysis was carried out using the numpy (http://numpy.scipy.org) and scipy (http://www.scipy.org/) modules of python. R^2 was obtained by linear regression based on the stats.linregress function of the numpy module. In this paper R^2 is defined as the correlation coefficient.

1.3 Pseudogene identification pipeline

The *P. tricornutum* pseudogene identification pipeline is shown in Figure 1 and each step is described as follows. (i) A six-frame translation of the entire genome was searched against the functional protein data set of *P. tricornutum* using tBlastn (main parameters: default SEG low-complexity filter parameters (12 2.2 2.5), -1E -4, -F T, -M BLOSUM62) and the Biopython (http://biopython.org/wiki/Biopython) module was used to parse the Blast results. (ii) All matches with more than 40% identity were extracted. (iii) Repetitive sequences were removed and identical sequences were merged. When the overlap of adjacent sequences was over 30 bp, the sequences were arranged in descending order according to their importance (the higher the *E*-value the lower their importance) and matching objects of low importance were removed. Sequences fulfilling the following three criteria were considered as the same candidate pseudogene (a) $|q21–q12| \leqslant \max(20, 0.2\times L)$, (b) $c21–c12 \leqslant 3000$, and (c) $c22–c21\leqslant 60$ (these parameters are the same as those that were defined previously [33]). The longest *P. tricornutum* intron was 9880 bp and the shortest was 20 bp. In total, 99.3% of the introns were shorter than 3000 bp (the intron data are from the previously published annotated genome [32]). (iv) The coding and non-coding sequences of the matched sequences were obtained using the C# code. (v) Sequences with *E*-values > 10^{-10} were removed. (vi) The functional genes and potential function genes were removed from the list of matched sequences. (vii) Redundant sequences and sequences containing repetitive and low complexity regions were also removed. (viii) Candidate pseudogenes were classified according to the sizes of gaps that they contained (gap size was obtained by the pairwise alignment between the candidate pseudogenes and the coding regions of the corresponding functional genes). If candidate pseudogene contained gaps that were longer than 20 bp (the shortest intron was 20 bp), it was regarded as a candidate duplicated pseudogene that contained introns; if the gaps were shorter than 20 bp, then the sequence was processed as a candidate pseudogene. (ix) The introns of a candidate duplicated pseudogene were analyzed by RepeatMasker (http://www.repeatmasker.org/). If the introns were identified as repeated sequences, then they were designated to be non-intron sequences and the candidate duplicated pseudogenes containing these sequences were considered as “disrupted” processed pseudogenes. These “disrupted” pseudogenes were regarded as being produced by the insertion of repeated sequences that occurred during the evolution of the processed pseudogenes. (x) Candidate processed pseudogenes were divided into processed pseudogenes and pseudogene fragments depending upon whether or not they spanned > 70% of the length of the corresponding functional gene. (xi) Processed pseudogenes were divided into “true” processed pseudogenes or “putative” processed pseudogenes depending on whether or not they contained an ORF disruption (frameshift and/or mutation).

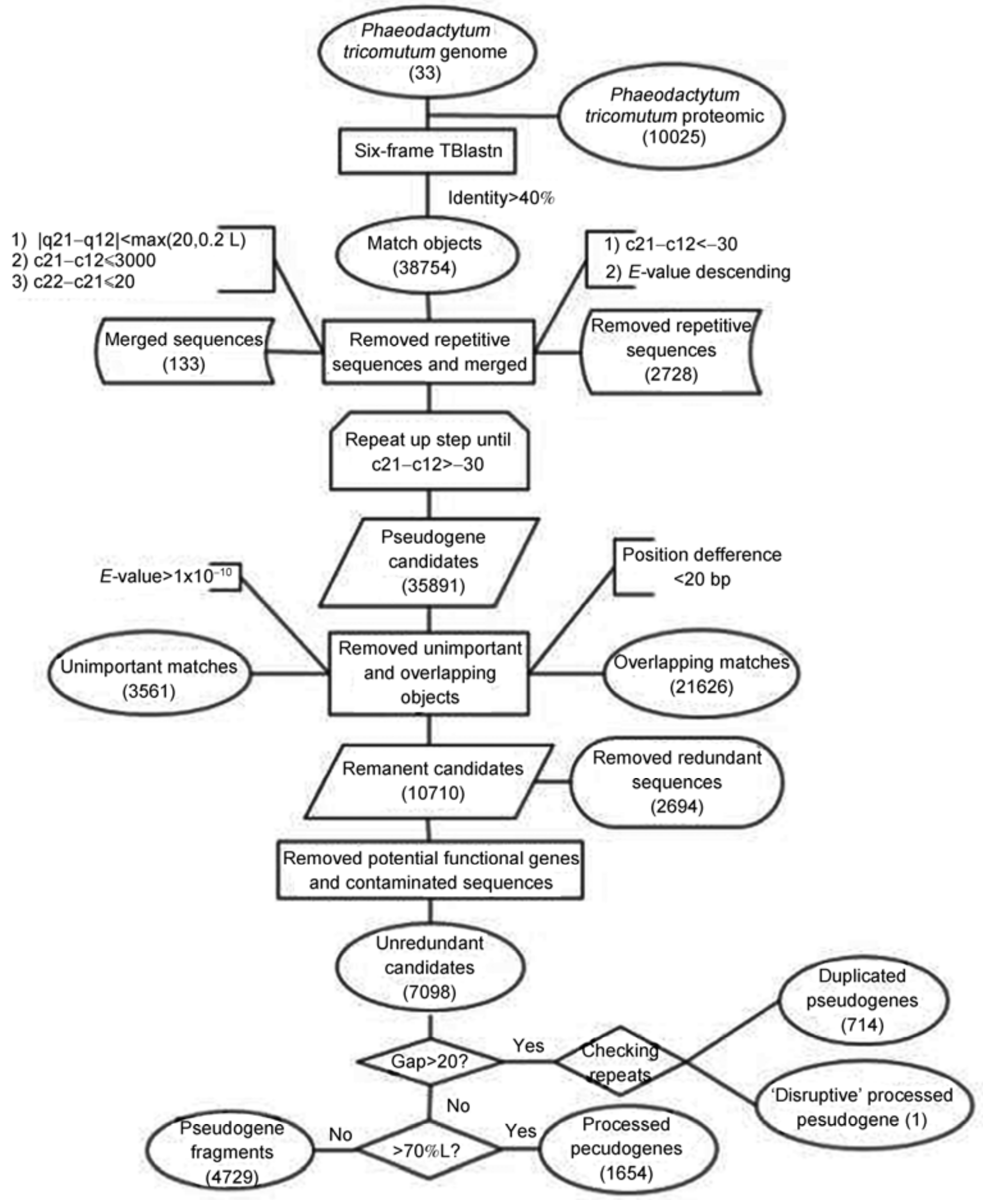
Phaeodactytum tricomutum genome (33)
Phaeodactytum tricomutum proteomic (10025)
Six-frame TBlastn
Identity>40%
1) |q21−q12|<max(20,0.2 L)
2) c21−c12≤3000
3) c22−c21≤20
Match objects (38754)
1) c21−c12<−30
2) *E*-value descending
Merged sequences (133)
Removed repetitive sequences and merged
Removed repetitive sequences (2728)
Repeat up step until c21−c12>−30
Pseudogene candidates (35891)
E-value>1x10^-10
Position defference <20 bp
Unimportant matches (3561)
Removed unimportant and overlapping objects
Overlapping matches (21626)
Remanent candidates (10710)
Removed redundant sequences (2694)
Removed potential functional genes and contaminated sequences
Unredundant candidates (7098)
Duplicated pseudogenes (714)
Gap>20?
Yes
Checking repeats
'Disruptive' processed pesudogene (1)
No
Pseudogene fragments (4729)
No
>70%L?
Yes
Processed pecudogenes (1654)

Figure 1 *P. tricornutum* pseudogene identification pipeline. The parameters are the same as those that have been defined previously [19]. Schematic graph showing the considerations used in merging two adjacent matches M1 and M2. (c11, c12) and (c21, c22) are chromosomal coordinates for M1 and M2; (q11, q12) and (q21, q22) are the corresponding matching regions on the query functional protein [33].

2 Results and discussion

2.1 Total number and distribution of pseudogenes in *P. tricornutum*

As shown in Figure 2, 1654 processed pseudogenes (23.3% of the total), 714 non-processed pseudogenes (10.1% of the total) and 4729 pseudogene fragments (66.6% of the total) were obtained by bioinformatics analysis. The processed pseudogenes were divided into three categories: "true" processed pseudogenes, "putative" processed pseudogenes, and "disruptive" processed pseudogenes (Table 1). There were less numbers of "true" processed pseudogenes than there were "putative"; there was only one "disruptive" processed pseudogenes, indicating that there were no insertions of repeated sequences in most of the ORFs of the processed pseudogenes. These results are very different from what has been reported for other species (such as human and rice [14,19]) in which "disruptive" pseudogenes are commonly found.

The distribution of processed pseudogenes, pseudogene fragments and functional genes in each *P. tricornutum* chromosome was analyzed. The results revealed that the distribution of pseudogenes on each of the chromosome was not equal (Figure 3). The largest numbers of "true" processed pseudogenes and "putative" pseudogenes (115 and 35

Table 1 Distribution of pseudogenes on each *P. tricornutum* chromosome

Chr	Size (kb)	GC content (%)	Functional gene	Processed pseudogene			Pseudogene fragment	Duplicated pseudogene	Density (per 0.1 Mb)
				True[a)]	Putative[b)]	Disrupted			
1	2535.40	47	1038	35	115		403	66	5.9
2	1497.95	47	611	6	73		152	31	5.3
3	1460.05	48	552	22	69		333	45	6.2
4	1360.15	48	538	12	45		177	27	4.2
5	1098.05	48	430	4	58		168	21	5.6
6	1035.08	49	445	6	54		126	24	5.8
7	1029.02	46	389	9	45	1	133	23	5.2
8	1007.77	48	401	4	33		154	13	3.7
9	1002.81	46	379	13			168	33	1.3
10	976.49	48	377	4	33		124	10	3.8
11	945.03	48	367	16	48		167	19	6.8
12	901.85	47	351	25	64		266	39	9.9
13	887.52	48	357	26(21)[c)]	68(1)[c)]		201(17)[c)]	32	10.6
14	829.36	48	316	14	29		112	24	5.2
15	814.91	49	321	8	28		115	30	4.4
16	764.23	48	291	29	44		183	17	9.6
17	703.94	49	260	15	33		195	36	6.8
18	702.47	49	259	11	63		235	14	10.5
19	690.43	49	258	7	28		83	22	5.1
20	683.01	49	260	11	41		129	30	7.6
21	662.22	49	254	3	34		86	14	5.6
22	591.34	48	236	5	20		83	15	4.2
23	512.49	49	192	0	24		109	6	4.7
24	511.74	48	158	4	19		139	18	4.5
25	497.27	49	182	13	36		111	20	9.9
26	441.23	47	141	11	38		62	13	11.1
27	404.30	48	147	9	21		93	16	7.4
28	387.58	47	167	4	7		50	8	2.8
29	384.26	41	96	7	24		122	15	8.1
30	317.21	49	104	4	13		74	5	5.4
31	258.24	48	78	12	40		101	18	20.1
32	157.05	40	44	1	8		32	8	5.7
33	87.97	45	23	3	10		43	2	14.8
Total	24612623	48	10022	353	1301	1	4729	714	

a) True processed pseudogenes contain frameshift disruption; b) putative processed pseudogenes contain no frameshift disruption; c) the numbers in brackets represent the number of pseudogenes containing a polyA tail.

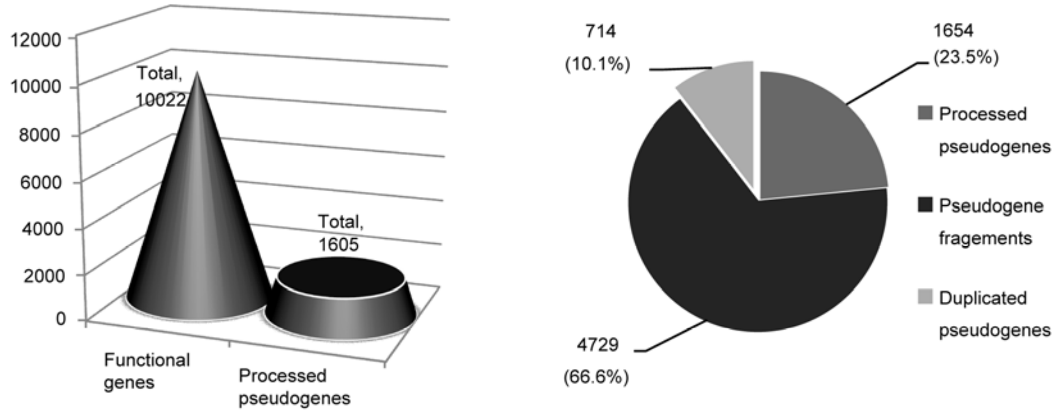

Figure 2 Statistics for the processed pseudogenes, pseudogene fragments, duplicated pseudogenes and functional genes.

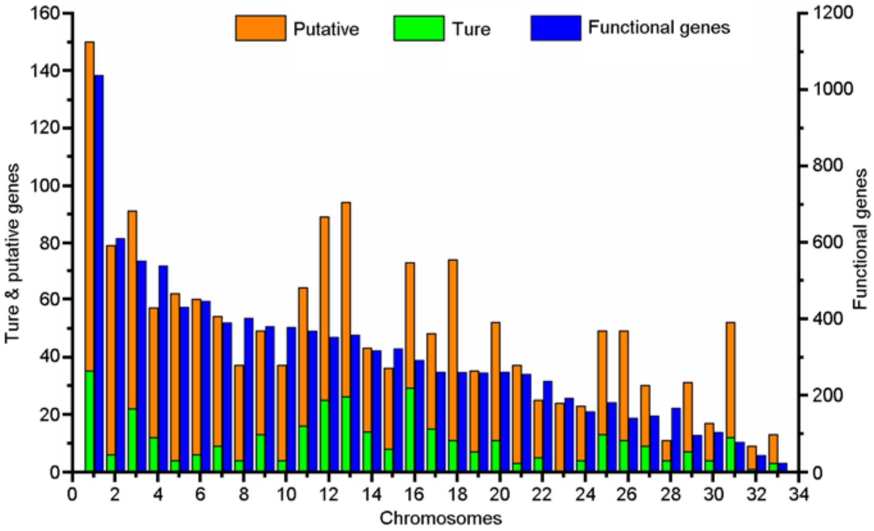

Figure 3 Distribution of "true" processed pseudogenes, "putative" processed pseudogenes, and functional genes across the different *P. tricornutum* chromosomes.

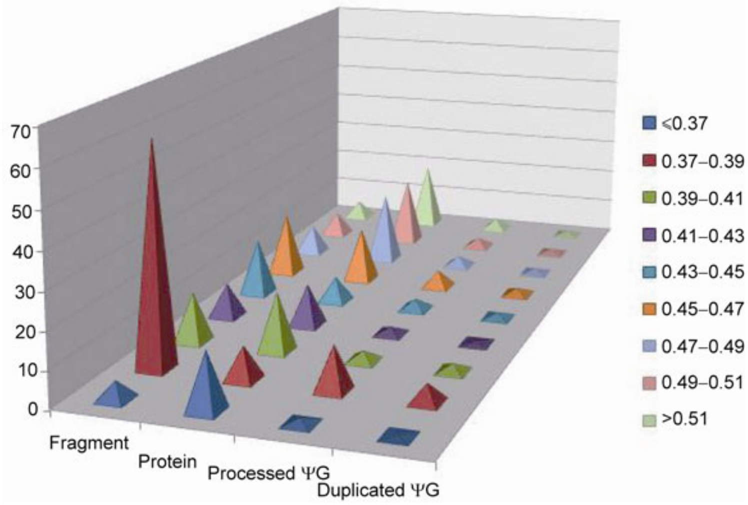

Figure 4 Distribution of pseudogene fragments, functional genes, processed pseudogenes, and duplicated pseudogenes related to the GC content of the *P. tricornutum* genome. The *X*-axis represents the sequence segments, the *Y*-axis represents the sequence densities (per 50 kb), and the *Z*-axis represents the GC content. The genome sequences were divided into 50 kb long segments and the number of pseudogene fragments, functional genes, processed pseudogenes, and duplicated pseudogenes in each segment were calculated and related to its GC content.

respectively) were on chromosome 1, and no "true" processed pseudogenes were found on chromosome 23. The density distribution of the processed pseudogenes and pseudogene fragments in the chromosomes were very similar; the distribution of pseudogene fragments and processed pseudogenes in regions with different GC content were also very similar (Figures S1 and 4). These results suggested that there was high correlation between the processed pseudogenes and the pseudogene fragments. Therefore, the selection of only the processed pseudogenes for further study should have no effect on the results of the analyses in this study.

Only the "true" and "putative" processed pseudogenes and the pseudogene fragments on chromosome 13 retained traces of the poly(A) tail and each of the poly(A) fragments were very short, about 15 bp (Table 1). It has been suggested that, over time, the remaining poly(A) fragments in the processed pseudogenes will slowly degrade or even disappear [33,34]. Thus, some of the pseudogenes in chromosome 13 of *P. tricornutum* might have emerged later than the pseudogenes in other chromosomes.

2.2 Correlation between chromosome length and the number of processed pseudogenes

As shown in Figures 5 and S2, there is a linear relationship between the number of processed pseudogenes or functional genes in a chromosome and the length of chromosomes (R^2 =0.61 and 0.99 respectively); that is, the longer the chromosome the greater the number of processed pseudogenes and functional genes that it contains. For example, there were 150 processed pseudogenes and 1038 functional genes in chromosome 1, the longest *P. tricornutum* chromosome, while in chromosome 33, the shortest of the chromosomes, there were only 13 processed pseudogenes and 23 functional genes. The longer chromosome sequences may present more opportunity for the insertion by retrotransposition of mRNA. This conclusion is consistent with the results of Zhang et al. [19] who analyzed pseudogenes in the human genome.

2.3 Correlation between the number of the processed pseudogenes and the GC content of the genome

The GC content of each chromosome was calculated and then the *P. tricornutum* genome sequence was cut into 50 kb-long non-overlapping segments and the GC content and number of pseudogenes in each segment was counted. For pseudogenes that were located between two adjacent segments, the segment that contained the longer part of the pseudogene was taken to be the segment that contained that pseudogene. In addition, the distribution of the 1654 processed pseudogenes, 4729 pseudogene fragments, functional protein and duplicated pseudogenes in each segment was analyzed. The results showed that the GC content on each *P. tricornutum* chromosome was evenly distributed, and was mainly between 45% and 49% (Figure S3 and Table 1). The results in Figure 4 indicate that the distribution of the functional proteins on the genome was fairly uniform, and consequently had no correlation with the genome internal GC content; on the other hand, pseudogene fragments, processed pseudogenes and duplicated pseudogenes tended to be enriched in areas of lower GC content (37%–41%) (Figure 4), possibly because of a negative selection pressure that might have confronted the pseudogenes [19]. However, as the data in Figure S4 show, there was no correlation between the average GC content of a chromosome and the density of the pseudogenes (per 0.1 Mb) that it contained. The results shown in Figure 5 revealed that the length of the chromosome sequence did play a dominant role in its pseudogene content, indicating that in local regions of *P. tricornutum* genome, pseudogenes tended to occur in areas with lower GC content possibly as the result of negative selection pressure [19].

2.4 Completeness of processed pseudogenes and their identity with corresponding functional gene sequences affects the number of pseudogenes

The ratio of the lengths of the pseudogene and the corresponding functional gene was important in deciding the completeness of the pseudogene sequences. If the ratio was greater than 70%, the pseudogene was considered to be a processed pseudogene; if not, it was classified as a pseudogene fragment [19]. The data in Figure 6 showed that most of the pseudogenes were almost complete after the retrotransposition; the average completeness of the “true” processed pseudogenes was 95% and for the “putative” processed pseudogenes was 92%. Of all the processed pseudogene sequences, 1000 (61%) were more than 90% complete (Table S1).

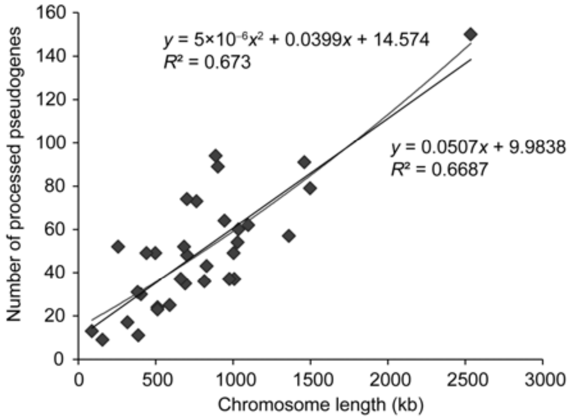

Figure 5 Correlation between chromosome length and the number of processed pseudogenes.

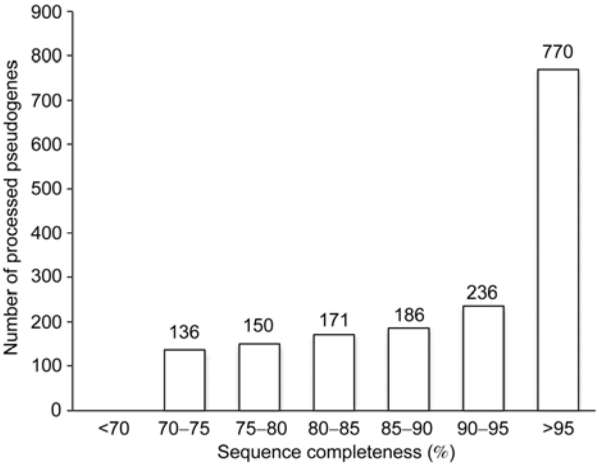

Figure 6 Correlation between the completeness of the pseudogene sequences and the number of processed pseudogenes in the *P. tricornutum* genome.

The pseudogenes had identities that were either less than 50% or greater than 90% with their corresponding functional gene (Figure 7); this finding is very different from what was reported for the processed pseudogenes in the human genome where the number of pseudogenes tended to increase with increasing identity [19]. Figure S5 shows the relationship between the identities of the *P. tricornutum* pseudogenes and their completeness. It was found that the higher the identity of the pseudogene with the corresponding functional gene, the higher the sequence completeness (R^2=0.72). It has been suggested that newly formed pseudogenes have a higher level of sequence completeness and identity compared to those that emerged earlier [19]. Therefore, the *P. tricornutum* pseudogenes with identities

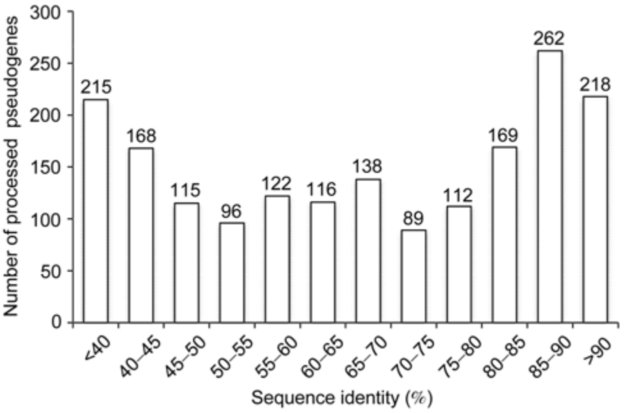

Figure 7 Correlation between the pseudogenes percentage identities with the corresponding functional genes and the number of processed pseudogenes.

greater than 90% may have formed more recently than the pseudogenes with identities less than 50%; these pseudogenes may have formed much earlier. This suggestion implies that there may have been two active periods during which the pseudogenes of *P. tricornutum* formed. During the intervening period, the environment in which *P. tricornutum* lived may have experienced significant changes.

2.5 Relationship between the number of processed pseudogenes or pseudogene fragments and the length of the corresponding functional genes

When the number of processed pseudogenes within each functional gene was calculated, it was found that most of the functional genes (9101) had no corresponding processed pseudogenes; 732 functional genes had one corresponding processed pseudogene and a very small number (77, about 0.8%) had more than three corresponding processed pseudogenes. A few functional genes (23) had more than 10 corresponding processed pseudogenes. Similar results were found for the number of pseudogene fragments that corresponded to functional genes. This trend can be illustrated by plotting the number of pseudogene fragments against the logarithm of the corresponding number of functional proteins (Figure 8).

Table S2 lists the corresponding homologous proteins that had more than 10 processed pseudogenes. Because the annotation of the *P. tricornutum* genome is incomplete, most of these homologous proteins were annotated as "predicted protein"; this made it difficult to functionally classify the homologous proteins that had the largest numbers of pseudogenes.

To analyze the relationship between the lengths of the functional proteins and the number of corresponding processed pseudogenes, the numbers of processed pseudogenes and pseudogene fragments in consecutive 200 amino acid long windows (with no overlap between two consecutive window positions) were calculated. The results showed that the number of processed pseudogenes and pseudogene fragments tended to decrease as the length of the functional genes increased (Figure S6). Thus, it would appear that the shorter mRNAs were more likely to be retrotransposed and to form processed pseudogenes or pseudogene fragments; the longer functional proteins, certainly their corresponding mRNAs, were less likely to be inserted into genomes by retrotransposition. Even if they did get inserted, they would be lost during evolution of the genome because of their greater influence on the structure and function of the genome. It has been reported that when mRNA sequences are longer than a specific value, they can no longer be retrotransposed in genomes [35].

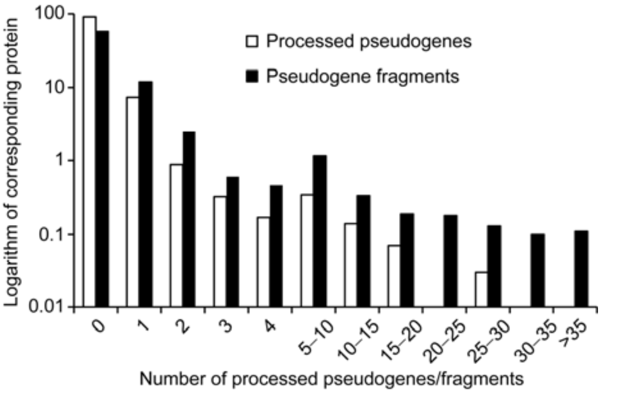

Figure 8 Number of proteins that have associated homologous processed pseudogenes or pseudogene fragments.

2.6 The distribution of pseudogenes and pseudogene fragments within the chromosomes

The distribution of the functional genes, pseudogenes and pseudogene fragments within each of the *P. tricornutum* chromosome was calculated (Figure S7). The results showed that the distribution of functional genes in each chromosome was more or less uniform, while no such rule could be applied to the distribution of the pseudogenes and pseudogene fragments in each chromosome.

3 Conclusions

Because the *P. tricornutum* genome was found to contain many functional genes derived from animals, plants and bacteria, the *P. tricornutum* genome has been reported to be more like an animal genome than a plant genome [32]. The method that has been used to identify pseudogenes in the human genome was optimized according to the features of *P. tricornutum* genome. Thus the method described in this paper was appropriately used to identify and analyze pseudogenes in the *P. tricornutum* genome. A total of 1654 processed pseudogenes, 714 duplicated pseudogenes, and 4729 pseudogene fragments that proved to be accurate and authentic were identified. As far as we know, this is the first pseudogene study on algae. The results should offer insights into the structure and evolution of the *P. tricornutum* genome.

This work was supported by the National Basic Projects for Science and Technology (SQ2012FY4910019-1), the Nonprofit Research Project for the State Oceanic Administration (201105008-2) and the National Natural Science Foundation of China (41176137).

1 Vanin E F. Processed pseudogenes: Characteristics and evolution. Annu Rev Genet, 1985, 19: 253–272

2 Mighell A J, Smith N R, Robinson P A, et al. Vertebrate pseudogenes. Febs Lett, 2000, 468: 109–114

3 Nowak M A, Boerlijst M C, Cooke J, et al. Evolution of genetic redundancy. Nature, 1997, 388: 167–171

4 Tachida H, Kuboyama T. Evolution of multigene families by gene duplication: A haploid model. Genetics, 1998, 149: 2147–2158

5 Force A, Lynch M, Pickett F B, et al. Preservation of duplicate genes by complementary, degenerative mutations. Genetics, 1999, 151: 1531–1545

6 Mccarrey J R, Thomas K. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature, 1987, 326: 501–505

7 Glusman G, Yanai I, Rubin I, et al. The complete human olfactory subgenome. Genome Res, 2001, 11: 685–702

8 Balakirev E S, Ayala F J. Pseudogenes: Are they "Junk" or functional DNA? Annu Rev Genet, 2003, 37: 123–151

9 Hirotsune S, Yoshida N, Chen A, et al. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature, 2003, 423: 91–96

10 Svensson O, Arvestad L, Lagergren J. Genome-wide survey for biologically functional pseudogenes. PLoS Comp Biol, 2006, 2: e46

11 Lerat E, Ochman H. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res, 2005, 33: 3125–3132

12 Harrison P, Kumar A, Lan N, et al. A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution1. J Mol Biol, 2002, 316: 409–419

13 Benovoy D, Drouin G. Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome. J Mol Evol, 2006, 62: 511–522

14 Nelson D R, Schuler M A, Paquette S M, et al. Comparative genomics of rice and *Arabidopsis*. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. Plant Physiol, 2004, 135: 756–772

15 Fran Oise T N, Shu O, Robin B C. Identification and characterization of pseudogenes in the rice gene complement. BMC Genomics, 2009, 10: 317

16 Harrison P M, Echols N, Gerstein M B. Digging for dead genes: An analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res, 2001, 29: 818–830

17 Harrison P M, Milburn D, Zhang Z, et al. Identification of pseudogenes in the *Drosophila melanogaster* genome. Nucleic Acids Res, 2003, 31: 1033–1037

18 Khelifi A, Duret L, Mouchiroud D. HOPPSIGEN: A database of human and mouse processed pseudogenes. Nucleic Acids Res, 2005, 33(Database issue): D59–D66

19 Zhang Z, Harrison P M, Liu Y, et al. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res, 2003, 13: 2541–2558

20 Torrents D, Suyama M, Zdobnov E, et al. A genome-wide survey of human pseudogenes. Genome Res, 2003, 13: 2559–2567

21 Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860–921

22 Yeh R F, Lim L P, Burge C B. Computational inference of homologous gene structures in the human genome. Genome Res, 2001, 11: 803–816

23 Field C, Behrenfeld M, Randerson J, et al. Primary production of the biosphere: Integrating terrestrial and oceanic components. Science, 1998, 281: 237

24 Falkowski P, Barber R, Smetacek V. Biogeochemical controls and feedbacks on ocean primary production. Science, 1998, 281: 200

25 Treguer P, Nelson D, Van Bennekom A, et al. The silica balance in the world ocean: A reestimate. Science, 1995, 268: 375

26 Werner D. The Biology of Diatoms: Silicate Metabolis. Berkeley: University of California Press, 1977. 149

27 Gibbs S. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. Ann Ny Acad Sci, 1981, 361: 193–208

28 Delwiche C, Palmer J. The origin of plastids and their spread via secondary symbiosis. Plant System Evol Suppl, 1997, 11: 53–86

29 Medlin L K, Kooistra W, Schmid A M M. The origin and early evolution of the diatoms: Fossil, molecular and biogeographical approaches. Cracow, Poland: W. Szafer Institute of Botany, Polish Academy of Sciences Press, 2000. 13–35

30 Mcfadden G, Van Dooren G. Evolution: Red algal genome affirms a common origin of all plastids. Curr Biol, 2004, 14: R514–R516

31 Nisbet R, Kilian O, Mcfadden G. Diatom genomics: Genetic acquisitions and mergers. Curr Biol, 2004, 14: 1048–1050

32 Bowler C, Allen A E, Badger J H, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature, 2008, 456: 239–244

33 Zhang Z, Harrison P, Gerstein M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res, 2002, 12: 1466–1482

34 Li W H, Gojobori T, Nei M. Pseudogenes as a paradigm of neutral evolution. Nature, 1981, 292: 237–239

35 Gon Alves I, Duret L, Mouchiroud D. Nature and structure of human genes that generate retropseudogenes. Genome Res, 2000, 10: 672–678