Expression Divergence of Duplicate Genes in the Protein Kinase Superfamily in Pacific Oyster

Libertas Academica
FREEDOM TO RESEARCH

Dahai Gao[1], Dennis C. Ko[2], Xinmin Tian[3], Guang Yang[4] and Liuyang Wang[2]

[1]Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, Shandong, China. [2]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC, USA. [3]Tropical Crops Genetic Resources Research Institute, Chinese Academy of Tropical Agricultural Sciences, Danzhou, Hainan, China. [4]Key Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, Shandong, China.

Supplementary Issue: RNA: An Expanding View of Function and Evolution

ABSTRACT: Gene duplication has been proposed to serve as the engine of evolutionary innovation. It is well recognized that eukaryotic genomes contain a large number of duplicated genes that evolve new functions or expression patterns. However, in mollusks, the evolutionary mechanisms underlying the divergence and the functional maintenance of duplicate genes remain little understood. In the present study, we performed a comprehensive analysis of duplicate genes in the protein kinase superfamily using whole genome and transcriptome data for the Pacific oyster. A total of 64 duplicated gene pairs were identified based on a phylogenetic approach and the reciprocal best BLAST method. By analyzing gene expression from RNA-seq data from 69 different developmental and stimuli-induced conditions (nine tissues, 38 developmental stages, eight dry treatments, seven heat treatments, and seven salty treatments), we found that expression patterns were significantly correlated for a number of duplicate gene pairs, suggesting the conservation of regulatory mechanisms following divergence. Our analysis also identified a subset of duplicate gene pairs with very high expression divergence, indicating that these gene pairs may have been subjected to transcriptional subfunctionalization or neofunctionalization after the initial duplication events. Further analysis revealed a significant correlation between expression and sequence divergence (as revealed by synonymous or nonsynonymous substitution rates) under certain conditions. Taken together, these results provide evidence for duplicate gene sequence and expression divergence in the Pacific oyster, accompanying its adaptation to harsh environments. Our results provide new insights into the evolution of duplicate genes and their expression levels in the Pacific oyster.

KEYWORDS: duplicate genes, Pacific oyster, RNA-seq, protein kinase superfamily

SUPPLEMENT: RNA: An Expanding View of Function and Evolution

CITATION: Gao et al. Expression Divergence of Duplicate Genes in the Protein Kinase Superfamily in Pacific Oyster. *Evolutionary Bioinformatics* 2015:11(S1) 57–65 doi: 10.4137/EBO.S30230.

TYPE: Original Research

RECEIVED: June 05, 2015. **RESUBMITTED:** July 29, 2015. **ACCEPTED FOR PUBLICATION:** August 03, 2015.

ACADEMIC EDITOR: Dr Jike Cui, Associate Editor

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 554 words, excluding any confidential comments to the academic editor.

FUNDING: This research was supported by the grants from National Science Foundation of China (No. 31402337) to D.G., and D.C.K. and L.W. were supported by Duke School of Medicine Whitehead Scholarship. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: wallacewly@gmail.com or liuyang.wang@duke.edu

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Gene duplication plays key roles in organismal evolution.[1,2] Duplicate genes initially have identical sequences but diverge in regulatory and coding regions during subsequent evolution. Divergence in regulatory regions could result in changes in expression levels, whereas changes in coding regions may lead to the acquisition of new functions.[3–5] The rapid development of next-generation sequencing technology in the past decade provides unique opportunities to study the general pattern from the whole genome and expression level. Indeed, several studies have attempted to characterize the correlation between expression patterns and genomic divergence for duplicate genes in human being, yeast, *Arabidopsis*, and cow.[5–9] Therefore, a general picture of the patterns of expression divergence in the evolution of duplicated genes is emerging. For instance, a positive correlation between synonymous sequence divergence and expression divergence was reported for human and yeast duplicate genes.[10] Likewise, the analysis of bovine duplicate genes also revealed that expression changes were correlated with sequence divergence.[11] On the other hand, unambiguous evidence of weak correlation between synonymous sequence divergence and expression divergence has been found in a case study of *Arabidopsis* duplicate genes.[12] Owing to these inconsistent findings within this limited number of species, there is still a need to characterize the relationship between expression and sequence divergence, especially in the nonmodel organism.

The Pacific oyster *Crassostrea gigas* is a representative species of phylum Mollusca, belonging to a large taxonomic group of protostomes and the group of marine animals with the largest number of identified species. Despite the species richness of this phylum, the genomes of Mollusca have only recently been examined. The whole-genome sequence and various developmental and stress-response RNA-seq transcriptomes for Pacific oyster were released recently,[13,14] rendering this species more suitable for the study of the evolution of gene duplication. Thus, several studies have discussed the structural and expression divergence of some rapidly expanding immune gene families in this species.[15,16] However, the patterns of divergence of duplicate genes with roles outside of immune function have been largely ignored.

In the present study, we selected a set of genes with broad importance in cell signaling, the protein kinase superfamily, to analyze the evolutionary pattern between sequence and gene expression for duplicate genes. The protein kinase superfamily is one of the largest gene families in eukaryotes, comprising 2%–4% of all genes in human being and in several model species.[17] The protein kinases are well known for regulating the majority of cellular pathways, especially those involved in signal transduction.[18,19] Thus, studying the evolutionary history of protein kinases provides a window to the evolution of many organism's signaling pathways. Therefore, we undertook the present study to identify duplicated protein kinase genes in oyster and to characterize the pattern of divergence between sequence and expression. A total of 64 putative duplicate gene pairs were identified from 320 protein kinase family members. We first show that these duplicate genes have experienced stronger selective constraints. We then find unequal distributions of correlation coefficients between duplicate genes for expression patterns under each of the five different developmental and stress-induced conditions. Finally, we investigated the relationship between sequence divergence and expression divergence and found that positive correlation exists between sequence divergence and expression divergence in each of the four conditions, suggesting that sequence divergence may generally explain the expression divergence under those conditions.

Results and Discussion

Identification of duplicate genes from the protein kinase family. Protein kinases represent one of the largest gene families in eukaryotes, and an enormous number of members have been reported in the model species. For example, there are 516, 238, and 425 kinase genes in human being, fruit fly, and nematode, respectively.[19] In the present study, a total of 320 protein kinase family members are identified from the genome of Pacific oyster, which account for about 1.1% of all predicted genes.

Based on rigorous phylogenetic and reciprocal BLAST analyses, a total of 64 pairs are identified as putative duplicated gene pairs (Fig. 1 and Table 1). Those 128 genes are located on 111 scaffolds, indicating the scattered and wide distribution on the whole genome. Nine of the pairs are present on the same scaffold instead of being present on two different scaffolds, indicating a pattern of tandem duplication. However, we note that the frequency of tandem duplication in protein kinase family may be underestimated because of incomplete genome assembly and annotations.

Comparisons of duplicate gene pairs showed that 13 pairs (20.3%) have equal exon numbers, and 14 pairs (21.9%) have almost identical exon numbers (their exon number difference ≤2) (Table 1). In the remaining 37 pairs (57.8%), the exon–intron structures are divergent between duplicate paralogs, indicating that exon–intron structural divergence is a common occurrence in oyster protein kinase genes. By analyzing 612 pairs of sibling paralogs from *Arabidopsis* and rice, Xu et al.[7] demonstrated that exon–intron structural variation is prevalent in three gene families and the proposed three mechanisms, including exon/intron gain/loss, exonization/pseudoexonization, and insertion/deletion, might contribute to the formation of structural changes. Their findings suggest that such structural divergences have played a vital role during the evolution of duplicate genes, and our findings are consistent with these studies.

Sequence divergence between duplicate gene pairs. To estimate the sequence divergence between duplicate gene paralogs, we calculate the synonymous (K_s) and nonsynonymous (K_a) substitution rates of coding sequences for each gene pair. The synonymous substitution rate K_s can be recognized as a proxy of divergence time between duplicated genes. The distribution of K_s has two major peaks around 0.2 and 4.1 in the density plot (Fig. 2A), indicating that those gene pairs originated at two major different stages and differed by evolutionary time. In addition, more than a half (61%) of duplicate pairs have K_s larger than 4, indicating highly diverged sequences and relatively long evolutionary time. In contrast, 25% of duplicate pairs have K_s less than 1, representing recently duplicated genes and relatively little sequence divergence.

The ω (K_a/K_s) values reflect selection pressure during evolution. For all studied pairs, the ω values were lower than 1, suggesting that those pairs were all evolving under purifying selection with putative functional constraints (Fig. 2B). Moreover, in recent duplicated pairs, there are some gene pairs with ω values higher than 0.4, indicating that the evolutionary constraint might be relaxed in some degree. Those genes subjected to relaxed purifying selection may tend to accumulate more mutations, altering gene structure and expression. Intriguingly, in our recent study of the oyster TNF superfamily, we also found that recently originated duplicate genes were under purifying selection.[15]

Expression patterns of duplicate gene pairs. In order to characterize the expression divergence for all gene pairs, the RNA-seq data collected from 69 developmental and stress-induced RNA-seq datasets have been analyzed (expression values and detailed information are given in Supplementary File 3). The Pearson's correlation coefficient r was calculated to quantify correlation between duplicate genes at the level of expression.

We also identified two main clades based on expression data from the RNA-seq data, which displayed opposite patterns. Clade 1 (upper in Fig. 3A), which include over two-thirds of the pairs, exhibited large positive r values under the majority of expression conditions, suggesting consistent

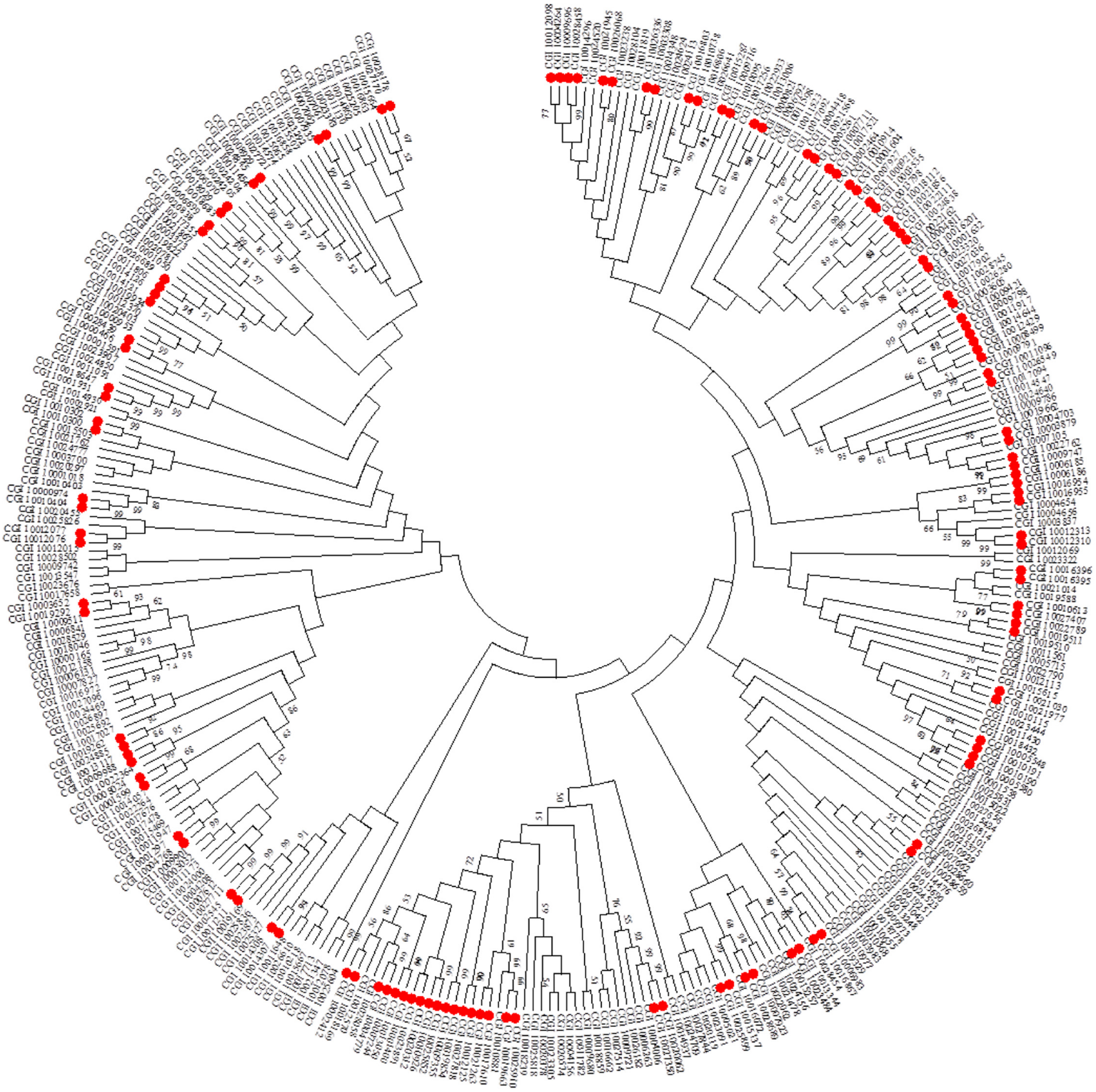

Figure 1. Phylogenetic relationship of protein kinases from Pacific oyster. NJ topology was represented and bootstrap values were shown for the clades with more than 50% support. The scale bar indicates the number of amino acid substitutions per site. The genes with red circles represent the identified duplicate paralogs.

expression patterns and similar functionality within each pair after their duplication. In contrast, Clade 2 (lower in Fig. 3A) showed a majority of negative *r* values, indicating that the paralogs in each pairs had divergent expression under most conditions.

For each condition of developmental and stress-induced transcriptomes, the median value of Pearson's *r* was positive. This indicated that most pairs have correlated expression patterns (Fig. 3B), suggesting that the genes in these pairs evolved under some functional constraint. Nonetheless, there is still a proportion of gene pairs also exhibiting negative *r* values, suggesting expression divergence within those gene pairs (Fig. 3B). Compared with the other four conditions of transcriptomes, the distribution of Pearson's *r* value under the heat condition was slightly lower, with a considerable proportion of negative values. These results suggest that those duplicated genes may have gained novel functions via subfunctionalization and/or neofunctionalization after their duplication. We hypothesize that these protein kinase genes have evolved to adapt to various stress environments or specialized developmental roles via expression divergence in oyster.

Positively correlated sequence divergence and expression divergence. The relationship between sequence divergence and expression divergence was investigated in all the five conditions of developmental and transcriptional transcriptomes (Fig. 4). We found significant negative

Table 1. Identified duplicate protein kinase gene pairs and related information.

PAIR NAME	GENE NAME	AMINO ACID LENGTH (aa)	SCAFFOLD	STRAND	EXON NUMBER
pair_01	CGI_10022762	291	scaffold443	–	8
	CGI_10009747	319	scaffold322	–	7
pair_02	CGI_10005548	317	scaffold268	–	6
	CGI_10010191	1058	scaffold930	–	13
pair_03	CGI_10028660	360	scaffold150	+	1
	CGI_10028659	325	scaffold150	–	4
pair_04	CGI_10006983	563	scaffold401	+	10
	CGI_10016867	989	scaffold1579	+	8
pair_05	CGI_10009421	672	scaffold116	–	25
	CGI_10009798	439	scaffold1560	–	13
pair_06	CGI_10011917	1281	scaffold1874	–	39
	CGI_10014644	1414	scaffold43964	–	21
pair_07	CGI_10012098	327	scaffold1195	+	10
	CGI_10004264	409	scaffold40612	+	2
pair_08	CGI_10009696	701	scaffold372	–	16
	CGI_10028458	792	scaffold102	–	7
pair_09	CGI_10021945	442	scaffold1086	–	11
	CGI_10026068	488	scaffold1174	+	18
pair_10	CGI_10026336	486	scaffold678	+	12
	CGI_10003308	360	scaffold39368	+	10
pair_11	CGI_10016803	1961	scaffold556	–	25
	CGI_10010738	350	scaffold954	+	9
pair_12	CGI_10015287	936	scaffold44008	–	26
	CGI_10009716	504	scaffold1028	+	13
pair_13	CGI_10022933	545	scaffold950	–	11
	CGI_10013006	595	scaffold1164	–	13
pair_14	CGI_10004418	1209	scaffold201	+	32
	CGI_10021568	1755	scaffold237	–	41
pair_15	CGI_10007711	482	scaffold42776	+	14
	CGI_10017521	1106	scaffold120	+	6
pair_16	CGI_10010914	794	scaffold1288	+	21
	CGI_10001604	528	C35776	–	14
pair_17	CGI_10009216	338	scaffold1688	–	9
	CGI_10003535	354	scaffold39740	+	10
pair_18	CGI_10018112	760	scaffold396	+	16
	CGI_10021856	689	scaffold164	+	16
pair_19	CGI_10022111	491	scaffold109	–	16
	CGI_10024838	468	scaffold492	–	13
pair_20	CGI_10016201	392	scaffold324	+	9
	CGI_10001632	444	scaffold277	+	9
pair_21	CGI_10028745	368	scaffold1009	+	8
	CGI_10026280	361	scaffold1836	–	11
pair_22	CGI_10014307	499	scaffold737	+	2
	CGI_10014308	784	scaffold737	+	2

(*Continued*)

Table 1. (*Continued*)

PAIR NAME	GENE NAME	AMINO ACID LENGTH (aa)	SCAFFOLD	STRAND	EXON NUMBER
pair_23	CGI_10011211	324	scaffold1157	+	1
	CGI_10002545	387	scaffold1795	–	8
pair_24	CGI_10004768	1373	scaffold1107	+	22
	CGI_10001297	894	C34444	–	12
pair_25	CGI_10001599	861	scaffold1453	+	18
	CGI_10008024	952	scaffold1277	+	21
pair_26	CGI_10009988	499	scaffold43366	–	13
	CGI_10013117	796	scaffold1252	+	19
pair_27	CGI_10024885	832	scaffold146	+	14
	CGI_10019262	615	scaffold506	+	10
pair_28	CGI_10019292	484	scaffold363	+	17
	CGI_10003652	993	scaffold1088	–	8
pair_29	CGI_10012076	1087	scaffold1492	–	8
	CGI_10012077	1082	scaffold1492	–	8
pair_30	CGI_10010404	1166	scaffold43446	–	16
	CGI_10000974	543	scaffold1496	–	6
pair_31	CGI_10010300	567	scaffold43426	–	2
	CGI_10010302	977	scaffold43426	–	2
pair_32	CGI_10001931	910	scaffold36398	+	17
	CGI_10018647	845	scaffold509	+	18
pair_33	CGI_10000466	269	C28760	+	5
	CGI_10028439	1336	scaffold102	–	21
pair_34	CGI_10014121	370	scaffold43932	–	8
	CGI_10014126	1593	scaffold43932	+	19
pair_35	CGI_10011806	539	scaffold43696	+	11
	CGI_10026689	677	scaffold53	–	10
pair_36	CGI_10020838	5054	scaffold1244	–	72
	CGI_10006699	1033	scaffold42366	+	25
pair_37	CGI_10018029	387	scaffold12	+	8
	CGI_10006070	862	scaffold1840	+	17
pair_38	CGI_10024845	327	scaffold492	+	6
	CGI_10008929	441	scaffold635	+	9
pair_39	CGI_10007062	831	scaffold1758	–	12
	CGI_10007061	994	scaffold1758	+	19
pair_40	CGI_10027170	770	scaffold1599	–	15
	CGI_10028178	1359	scaffold86	–	25
pair_41	CGI_10012429	2389	scaffold498	+	19
	CGI_10008499	325	scaffold43036	+	6
pair_42	CGI_10011096	516	scaffold340	–	14
	CGI_10026549	479	scaffold145	+	6
pair_43	CGI_10004703	474	scaffold1231	+	8
	CGI_10003879	403	scaffold40120	+	8
pair_44	CGI_10006185	585	scaffold1526	–	16
	CGI_10006186	730	scaffold1526	+	19

(*Continued*)

Table 1. (*Continued*)

PAIR NAME	GENE NAME	AMINO ACID LENGTH (aa)	SCAFFOLD	STRAND	EXON NUMBER
pair_45	CGI_10016954	667	scaffold117	+	15
	CGI_10016955	621	scaffold117	+	15
pair_46	CGI_10012313	720	scaffold477	–	18
	CGI_10012310	720	scaffold477	–	18
pair_47	CGI_10016396	252	scaffold594	+	3
	CGI_10016395	466	scaffold594	+	3
pair_48	CGI_10010613	594	scaffold43500	+	10
	CGI_10027407	432	scaffold1179	–	12
pair_49	CGI_10022789	283	scaffold443	+	7
	CGI_10019511	347	scaffold376	+	7
pair_50	CGI_10021030	1493	scaffold672	–	36
	CGI_10021977	774	scaffold1108	–	19
pair_51	CGI_10010190	562	scaffold930	–	10
	CGI_10005580	517	scaffold1708	+	10
pair_52	CGI_10007244	271	scaffold493	+	9
	CGI_10004779	599	scaffold1067	–	15
pair_53	CGI_10018169	273	scaffold459	+	7
	CGI_10002412	290	scaffold857	+	6
pair_54	CGI_10001400	485	scaffold34994	–	11
	CGI_10013050	331	scaffold43836	–	7
pair_55	CGI_10025852	784	scaffold1583	+	16
	CGI_10010926	384	scaffold1288	+	14
pair_56	CGI_10020312	344	scaffold522	+	9
	CGI_10023891	368	scaffold48	–	11
pair_57	CGI_10019854	401	scaffold1512	+	4
	CGI_10009355	1247	scaffold43208	+	10
pair_58	CGI_10012125	434	scaffold1890	+	13
	CGI_10027818	482	scaffold198	+	13
pair_59	CGI_10017610	475	scaffold1670	+	13
	CGI_10021263	584	scaffold157	+	13
pair_60	CGI_10025910	645	scaffold334	–	11
	CGI_10019663	357	scaffold1715	+	9
pair_61	CGI_10020062	555	scaffold258	+	1
	CGI_10027350	709	scaffold1179	–	14
pair_62	CGI_10013344	530	scaffold1894	–	11
	CGI_10023484	516	scaffold1258	–	13
pair_63	CGI_10007923	1373	scaffold42850	+	17
	CGI_10028689	1383	scaffold150	+	24
pair_64	CGI_10015137	208	scaffold1671	+	1
	CGI_10025899	600	scaffold733	+	2

correlation between transformed r' and K_a (or K_s) in four of the five conditions ($P < 0.05$, Fig. 4B–E). Interestingly, there is a significant positive correlation between expression divergence and sequence divergence among duplicate pairs under those conditions. However, for the set of transcriptomes comparing relative tissue abundance, the correlation was not statistically significant ($P = 0.201$ for K_a and $P = 0.436$ for K_s, Fig. 4A), suggesting less correlation

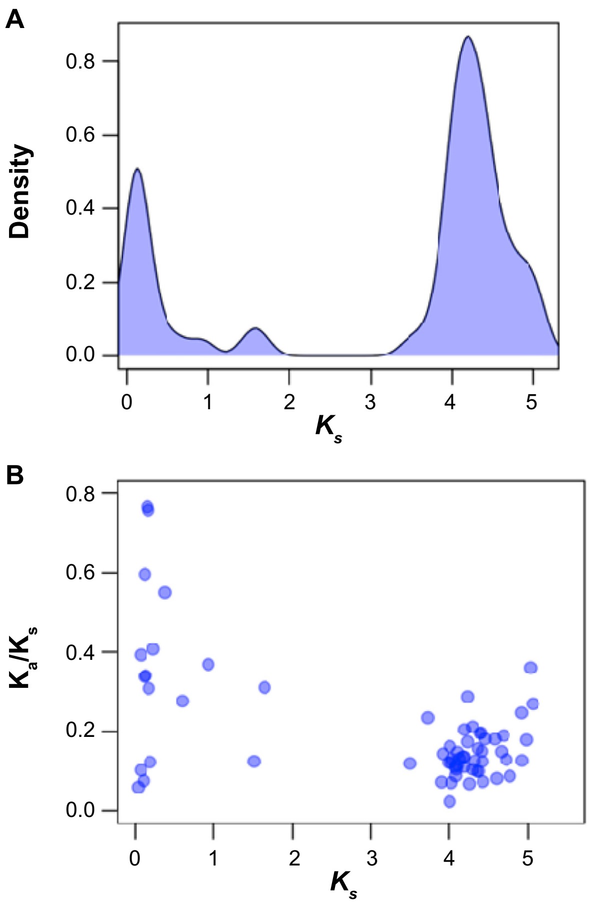

Figure 2. The sequence divergence between duplicate pairs. (**A**) The density distribution of synonymous rate (K_s) for all duplicate gene pairs. (**B**) The comparisons of K_a/K_s and K_s values, where K_s is a proxy of divergence time between duplicated genes.

between expression divergence and genome divergence. This pattern is mostly consistent with previous studies in yeast, human being, and cow.[8,9,11]

Gene duplications are widely present in eukaryotic genomes, providing increased opportunities for nonreciprocal recombination and allowing redundant genes to evolve new functions. However, the fate of duplicate genes is a widely discussed topic of genome evolution. Recently, the subfunctionalization and neofunctionalization models have been invoked to explain the retention of duplicate genes.[2,20] In this study, we found that most gene pairs exhibited consistent expression patterns and underwent purifying selection. Sequence and expression divergence were positively correlated under four conditions, consistent with the hypothesis of sequence divergence driving expression divergence. For the transcriptomes of tissue expression levels, we observed nonsignificant correlation between expression divergence and genome divergence, possibly because of genome divergence in noncoding regions not being reflective of the pattern seen in coding regions (as exemplified by K_a and K_s), a complicated expression divergence pattern, or the limited sample size used in our analysis. Therefore, we hypothesize that the functional constraints of protein kinase genes may contribute to the evolution of the duplicate paralogs in oyster. To sum up, our results provide insights into duplicate gene sequence and expression divergence in the Pacific oyster and may help to elucidate its adaptation to different environments and development processes. Our results may also help to understand the mechanisms for the retention of duplicate genes in other gene families in Pacific oyster.

Methods

Identification of duplicate gene pairs. The Pacific oyster genome sequences were downloaded from OysterDB (http://oysterdb.cn/home.html). The hidden Markov model (HMM) method was carried out to retrieve sequences containing a protein kinase domain (PF00069). The presence of a protein kinase domain (PF00229) was validated by SMART (http://smart.embl-heidelberg.de/) and Pfam (http://pfam.sanger.ac.uk/), and a total of 320 protein kinase sequences were identified from the Pacific oyster genome (see Supplementary File 1 for details). Phylogenetic analyses and reciprocal BLAST were applied to identify the duplicate gene pairs and relationships. First, phylogenetic analysis was carried out to identify the sequence pairs with close evolutionary relationships. We used the HMMalign program[21] to generate sequence alignments for protein kinase domain regions (provided as Supplementary File 2). The phylogenetic tree was reconstructed using the neighbor-joining (NJ) method from the MEGA 5.01 program[22] and a total of 75 paralog pairs (Fig. 1) were identified. Second, the sibling paralog relationship for each pair was further confirmed by reciprocal best BLAST hits analysis. The BLASTP program was used to compare each protein in identified pairs against all other proteins with the *E*-value cutoff of 1e-5. As a result, a total of 64 duplicate gene pairs were identified (Table 1), and their genome location and annotation were parsed from GFF files by a custom *Python* script.

Sequence alignment and divergence analysis. Protein sequences for each duplicated gene pair were first aligned by ClustalW.[23] Then, the PAL2NAL program was used to generate the codon alignment.[24] The synonymous (K_s) and nonsynonymous (K_a) substitution rates of coding sequences for duplicate pairs were calculated by the KaKs_Calculator[25] using the modified Yang–Nielsen algorithm (MYN).[26]

Expression profile analysis. The available expression values for *C. gigas* protein kinase genes under five varied condition and 69 RNA-seq transcriptome datasets (nine tissues, 38 developmental stages, eight dry treatments, seven heat treatments, and seven salty treatments) were obtained from the transcriptome dataset of oyster genome project (http://oysterdb.cn/home.html, Zhang et al.[13]). The Reads per Kilobase of exon Model per million mapped reads (RPKM) values were calculated to indicate the expression levels of each gene. The pairwise expression patterns for each gene pair under different conditions were visualized through heatmap in *R* version 2.13. Pearson's correlation coefficient *r* of

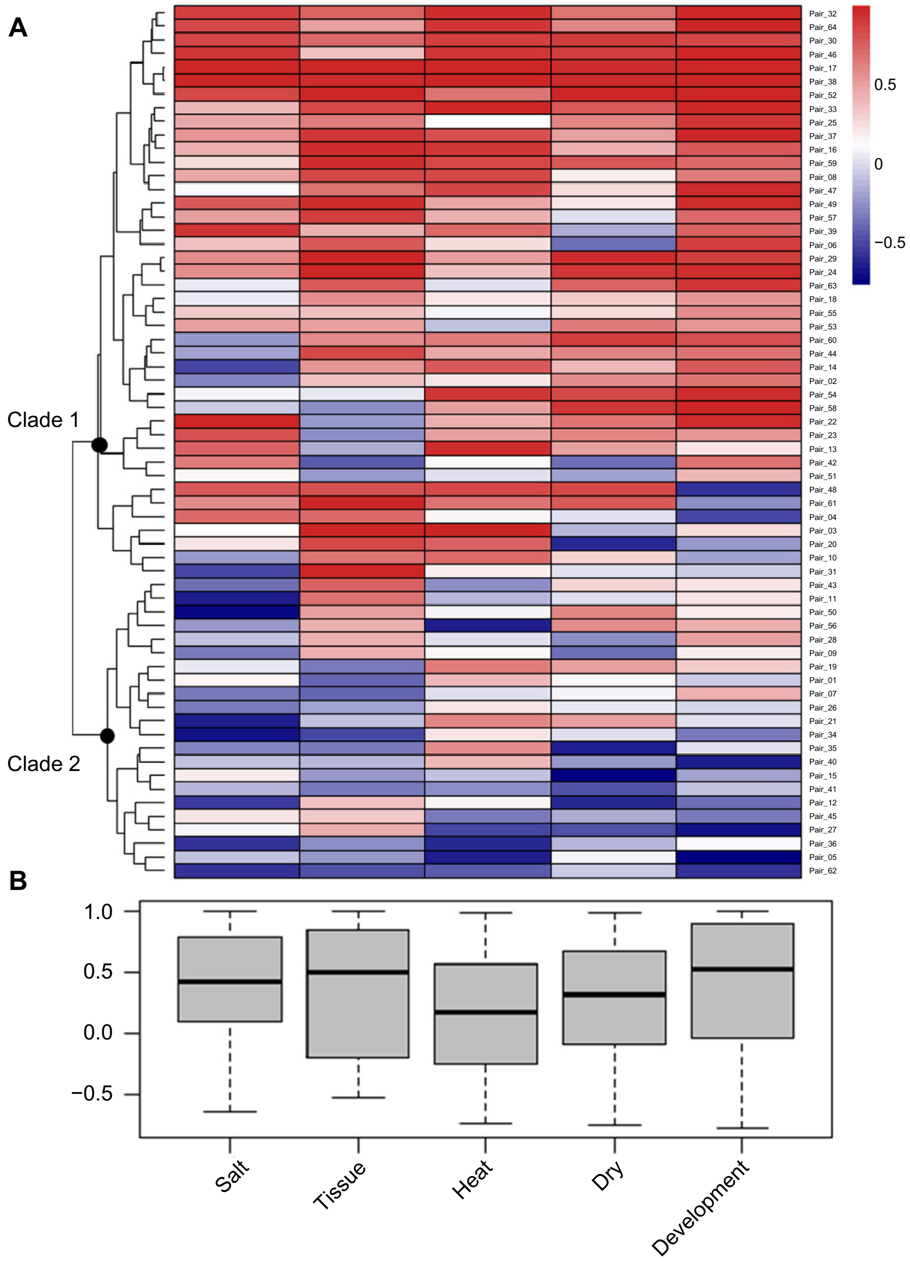

Figure 3. Expression patterns of duplicate gene pairs. (**A**) The heatmap was performed using Pearson's correlation coefficient of gene expression under five conditions, and red to blue blocks indicate high-to-low correlation levels. (**B**) The boxplot represents the distribution of correlation coefficient values for each expression condition.

expression level was calculated to measure the correlation among the duplicate gene pairs at the level of expression. Following previous studies, the Pearson's correlation coefficient r was further transformed to r' using equation

$$r' = \frac{\ln(1+r)}{1-r}$$

and the rescaled r' is more appropriate for linear regression analysis.[8,11] With the transformation, the negative regression coefficient between r' and K_s (or K_a) represents a positive relationship between expression level and K_s (or K_a).

Acknowledgments

We thank Dr. Benjamin Redelings for helpful suggestions and language editing. We also thank three anonymous reviewers for their valuable comments.

Author Contributions

Conceived and designed the experiments: DG, LW. Analyzed the data: DG, LW. Wrote the first draft of the manuscript: DG, DK, GY, LW. Contributed to the writing of the manuscript: DG, DK, XT, LW. Agreed with manuscript results and conclusions: DG, DK, XT, GY, LW. All the authors reviewed and approved the final manuscript.

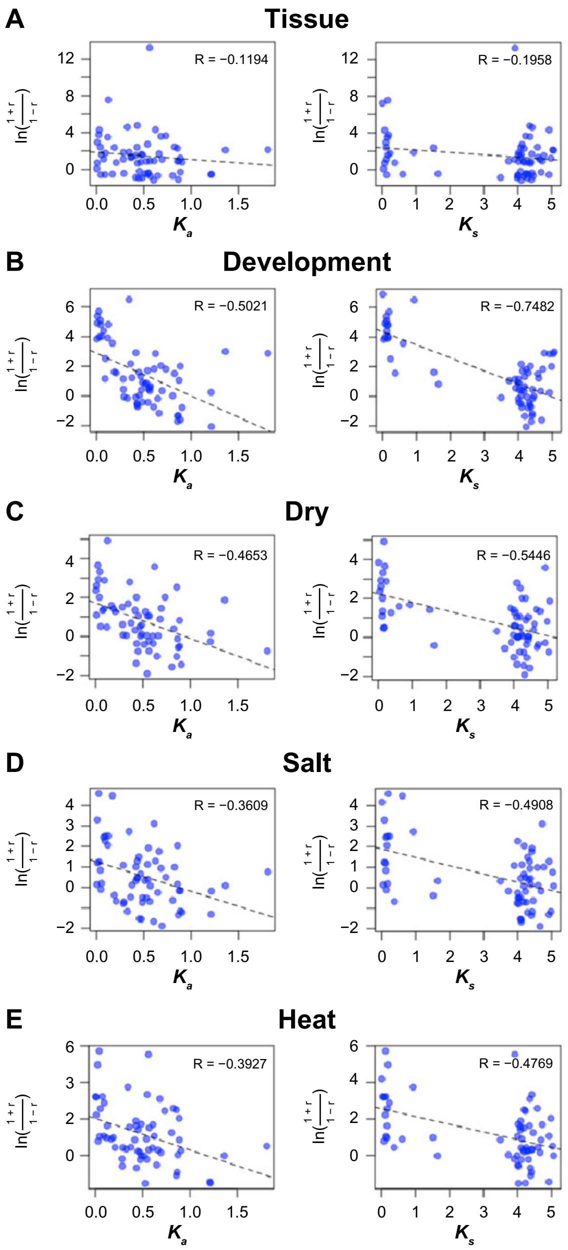

Figure 4. The relationship between the correlation coefficient (R) of gene expression and K_a (or K_s) in duplicate genes. (**A**) No correlation between $\frac{\ln(1+r)}{1-r}$ and K_a (or K_s) for tissue expression transcriptomes. (**B–E**) Negative correlations between $\frac{\ln(1+r)}{1-r}$ and K_a (or K_s) under developmental stages, dry treatments, salt treatments, and heat treatments, respectively. These imply positive correlation between sequence divergence and expression divergence because $1 - r$ can be regarded as expression divergence. Each point represents one gene pair.

Supplementary Material

Supplementary File 1. Full protein coding sequences of identified Pacific oyster protein kinase proteins genes (FASTA format).

Supplementary File 2. Amino acid alignment of identified Pacific oyster protein kinase proteins sequences (FASTA format).

Supplementary File 3. RPKM values for identified Pacific oyster protein kinase genes from five different developmental and stimuli-induced datasets (xlsx format).

REFERENCES

1. Ohno S. *Evolution by Gene Duplication*. London: George Alien & Unwin Ltd, Berlin, HD, New York: Springer-Verlag; 1970.
2. Zhang J. Evolution by gene duplication: an update. *Trends Ecol Evol*. 2003; 18:292–8.
3. Lynch M. Gene duplication and evolution. *Science*. 2002;297:945–7.
4. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*. 2010;11:97–108.
5. Wang Y, Wang X, Paterson AH. Genome and gene duplications and gene expression divergence: a view from plants. *Ann N Y Acad Sci*. 2012;1256:1–14.
6. Li Z, Zhang H, Ge S, Gu X, Gao G, Luo J. Expression pattern divergence of duplicated genes in rice. *BMC Bioinformatics*. 2009;10:S8.
7. Xu G, Guo C, Shan H, Kong H. Divergence of duplicate genes in exon-intron structure. *Proc Natl Acad Sci U S A*. 2012;109:1187–92.
8. Li WH, Yang J, Gu X. Expression divergence between duplicate genes. *Trends Genet*. 2005;21:602–7.
9. Makova KD, Li WH. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res*. 2003;13:1638–45.
10. Gu Z, Nicolae D, Lu HH, Li W-H. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet*. 2002;18:609–13.
11. Liao X, Bao H, Meng Y, Plastow G, Moore S, Stothard P. Sequence, structural and expression divergence of duplicate genes in the bovine genome. *PLoS One*. 2014;9:e102868.
12. Haberer G, Hindemitt T, Meyers BC, Mayer KF. Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of *Arabidopsis*. *Plant Physiol*. 2004;136:3009–22.
13. Zhang G, Fang X, Guo X, et al. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*. 2012;490:49–54.
14. Zhang L, Li L, Zhu Y, Zhang G, Guo X. Transcriptome analysis reveals a rich gene set related to innate immunity in the Eastern oyster (*Crassostrea virginica*). *Mar Biotechnol (NY)*. 2014;16:17–33.
15. Gao D, Qiu L, Gao Q, Hou Z, Wang L, Song L. Repertoire and evolution of TNF superfamily in *Crassostrea gigas*: implications for expansion and diversification of this superfamily in *Mollusca*. *Dev Comp Immunol*. 2015;51:251–60.
16. Zhang L, Li L, Guo X, Litman GW, Dishaw LJ, Zhang G. Massive expansion and functional divergence of innate immune genes in a protostome. *Sci Rep*. 2015;5:8693.
17. Manning G, Plowman GD, Hunter T, Sudarsanam S. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci*. 2002;27:514–20.
18. Shiu SH, Bleecker AB. Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in *Arabidopsis*. *Plant Physiol*. 2003;132:530–43.
19. Taylor SS, Kornev AP. Protein kinases: evolution of dynamic regulatory proteins. *Trends Biochem Sci*. 2011;36:65–77.
20. Nei M. *Mutation-Driven Evolution*. Oxford University Press; 2013. Oxford.
21. Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998;14:755–63.
22. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28:2731–9.
23. Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*. 2002;2:3.
24. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006;34:W609–12.
25. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics*. 2010;8:77–80.
26. Zhang Z, Li J, Yu J. Computing *Ka* and *Ks* with a consideration of unequal transitional substitutions. *BMC Evol Biol*. 2006;6:44.