Evolutionary and functional diversity of green fluorescent proteins in cephalochordates

Guang Li a, Qiu-Jin Zhang a,b, Jing Zhong a, Yi-Quan Wang a,*

a Key Laboratory of the Ministry of Education for Cell Biology and Tumor Cell Engineering, School of Life Sciences, Xiamen University, Xiamen, 361005, China
b College of Life Sciences, Fujian Normal University, Fuzhou, 350108, China

ARTICLE INFO

Article history:
Received 30 April 2009
Received in revised form 30 June 2009
Accepted 3 July 2009
Available online 15 July 2009

Keywords:
GFP
Amphioxus
Cephalochordate
Evolution

ABSTRACT

Green fluorescent protein (GFP) has been widely used as a molecular marker in modern biological research. Before the recent report of one GFP gene in Branchiostoma floridae, GFP family members were cloned only from other two groups of species: Cnidaria and Copepoda. Here we describe the complete GFP gene repertoire of B. floridae which includes 13 functional genes and 2 pseudogenes, representing the largest GFP family found so far. Coupling with nine other GFP sequences from another two species of genus Branchiostoma and the sequences from Cnidaria and Copepoda, we made a deep-level phylogenetic analysis for GFP genes in cephalochordates and found: 1) GFP genes have experienced a divergent evolution in cephalochordates; 2) all amphioxus GFP genes form four main clades on the tree which had diverged before the radiation of the last common ancestor of all extant cephalochordates; 3) GFP genes in amphioxus shared a common ancestor with that in Copepoda rather than being derived from horizontal gene transfer, which indicates that our ancestor was derived from a fluorescent organism and lost this ability after its separation from Cephalochordata, and also makes GFP a rare gene which has a rather unusual evolutionary path. In addition, we also provided evidence indicating that GFP genes have evolved divergent functions by specializing their expression profile, and different fluorescent spectra by changing their emission peaks. These findings spark two interesting issues: what are GFP in vivo functions in cephalochordates and why they are lost in other examined deuterostomes?

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

GFP protein is one of the most popular biomarkers applied in research on gene expression and protein localization. The first GFP gene was cloned from the luminescent jellyfish Aequorea victoria (Prasher et al., 1992) and after that many other GFP family members of divergent color emissions were characterized in several other species (Matz et al., 1999; Labas et al., 2002; Gurskaya et al., 2003; Shagin et al., 2004) which are derived from two metazoan groups: Cnidaria and Copepoda. However it was not until 2007 when a GFP gene was reported from a cephalochordate (amphioxus), Branchiostoma floridae (Deheyn et al., 2007), no GFP family members were found outside the above two groups of species, including the divergent model animals whose genome sequences are currently available. This widely scattered distribution of GFP genes in the animal kingdom makes their origin illusive. In this study, we characterized 23 new GFP genes (including two pseudogenes) from three species of genus Branchiostoma and made a systematical phylogenetic analysis for them. We also observed GFP proteins of different fluorescent spectra and provided evidence for their functional diversity at expression level by analyzing the promoter sequences and EST data.

Previous studies about GFPs mainly focused on how to extend their usefulness, however, much remains to be learned about biological function of these proteins in nature. Amphioxus, one of the vertebrates’ closest relatives, is an emerging model organism for studying the basic developmental mechanisms of chordate body plan and the evolutionary origin of vertebrate gene families (Holland et al., 2004). The diversity of amphioxus GFP gene shown here prohibits the possibility of GFP protein as a reporter in this species, but puts the issue about GFP in vivo functions on front line. Undoubtedly, amphioxus will be a perfect model to address this issue in the future.

2. Materials and methods

2.1. Characterization of GFP genes from three species of genus Branchiostoma

In this study, GFP genes were characterized from three amphioxus species: two Atlantic species (B. floridae, B. lanceolatum) and one Pacific species (B. japonicum, identified according to Xu et al. (2005) and Zhang et al. (2006)). The sequences of B. floridae were extracted from the genome database at JGI (http://genome.jgi-psf.org/) using...
the TBLASTN (Altschul et al., 1990) program with the previously reported GFP protein sequence (Deheyn et al., 2007) as query and cutoff E-value at 1e−5. The version I assembly used in this study contains sequences of two haplotypes. Therefore, if two alleles of one gene were found in the database, we selected the haplotype with more complete assembly for our analysis. We distinguished alleles from different gene copies as described by Li et al. (2007). mVISTA at http://genome.lbl.gov/vista/index.shtml was implemented to predict un-annotated genes with the previously reported GFP sequence as reference, and some of the predictions were verified by searching the EST database. The GFP sequences of B. lanceolatum were retrieved from NCBI (http://www.ncbi.nlm.nih.gov) using the BLASTP program with default settings. Twenty two B. lanceolatum GFP-related sequences were extracted. However, due to no genome data available it is hard to determine which of them are alleles and which of them are gene copies. To make it simple, we implemented the Sequencher 4.2 to cluster the above 22 sequences using assembly parameters: minimum match percentage = 95% and minimum overlap = 50 bp. The sequences of B. japonicum were cloned from cDNA templates using PCR strategy. To distinguish GFP genes in the above three amphioxus species, we named these genes by adding the first letter of their Latin names. For example, we added “Bf-” before the name of B. floricidum GFP gene.

2.2. B. floricidum GFP EST data treatment

We chose B. floricidum EST data at http://amphioxus.icob.sinica.edu.tw in our analysis because it was better compiled than that from NCBI in terms of sequencing directions and colony information. Each annotated amphioxus GFP sequence was employed to search against the database. A total of 472 GFP-related EST sequences were extracted, and among them, 376 are derived from 5′ sequencing. We applied a more complex strategy to cluster ESTs and the annotated GFP cDNAs, since several B. floricidum GFP genes are derived from very recent duplications. Firstly, all ESTs and B. floricidum GFP cDNA sequences were imported into the Sequencher 4.2 for cluster assembling. Sequences with a minimum overlap length of 100 bp and at least 95% identity were assembled into the same cluster. Then, because sequences in the same cluster diverge mainly in the UTR regions, we extracted the coding sequences from the clusters and some singletons, and then re-assembled them again under more strict conditions: minimum overlap length = 100 bp and minimum identity = 97%. For clusters or singletons which are still can’t assemble with the annotated GFP cDNAs, we adopted genomic sequence-based method to cluster them.

2.3. B. floricidum GFP gene promoter analysis by mVISTA

Bf-GFPs upstream intergenic regions and their intron/exon regions were aligned by mVISTA with more than 70% identity over 30 bp. We use such low constraint for two reasons. Firstly, these genes are derived from single ancestral gene by duplication and distributed in highly compact genome areas, so mVISTA will give reliable alignment. Secondly, most transcription factor binding sites are very small (~10 bp) but are highly conserved. The intergenic region was determined based on gene models predicted by JGI. Since repeat-masked genome sequence at JGI cannot be downloaded conveniently, we extracted the corresponding sequence from the UCSC (http://genome.ucsc.edu).

2.4. Evolutionary analysis

GFP protein sequences were aligned using CLUSTAL_X (Thompson et al., 1997), and the sequence matrix was applied to generate neighbor-joining tree (Saitou and Nei, 1987) using MEGA4 (Tamura et al., 2007) with protein Poisson distances (Nei and Kumar, 2000), and ML (maximum likelihood) tree using PHYML (Guindon et al., 2005) with WAG+I+G model. ProtTest was used to choose the best amino acid substitution model for PHYML analysis (Abascal et al., 2005). The reliability of each interior branch of the tree was assessed by bootstrapping with 100 replications (Felsenstein, 1985) and the following GFP sequences from Cnidaria were used as outgroups for
the NJ tree: AY485333 (phiYFP); AY485334 (anm1GFP1); AY485335 (anm1GFP2); AY485336 (anm2CP); X83960 (Av-GFP). In addition, we also included the GFP sequences from Copepoda in our dataset for calculating the divergence time between Cephalochordata and Copepoda: AY268072 (ppluGFP2); AY268073 (laesGFP); AY268074 (pmeaGFP1); AY268075 (pmeaGFP2); AY268076 (pdae1GFP). Names of GFP genes from Cnidaria and Copepoda are according to Shagin et al. (2004).

Fig. 2. Phylogeny tree of six representative cephalochordate species according to Kon et al. (2007) (a) and phylogenetic relationship of GFP genes (b). E. maldivensis, A. inferom and A. lucayanum on the phylogeny tree are abbreviations of Epigonichthys maldivensis, Asymmetron inferno and Asymmetron lucayanum respectively. The GFP gene tree was reconstructed by the neighbor-joining method with protein Poisson distances followed by linearization. Bl-GFP3 was not included in molecular clock analysis due to its extraordinarily slow evolutionary rate, but was put on the tree according to NJ topology constructed using all sequences. Bootstrap percentages over 50% are shown on interior branches. Five sequences from Cnidaria are used as outgroup. The divergence time at nodes marked with square black boxes was set as 112 Mya to calibrate the clock respectively. The divergence times between GFP genes from Cephalochordata and Copepoda, and between the four clades of Cephalochordate GFP genes are shown with the upper value being calibrated by GFP7 node and the lower by GFP9 node. The point indicated by the black dotted line is when the last common ancestor of all extant cephalochordates diverged (Nohara et al., 2005). GenBank identities for B. lanceolatum GFP genes are: BI-GFP1, EU482388; BI-GFP2, EU482390; BI-GFP3, EU482393; BI-GFP4, EU482400; BI-GFP5, EU482404; BI-GFP6, EU482405.
We performed branch-length test in LINTRDOS (Takezaki et al., 1995) to find and eliminate sequences which do not fit the molecular clock. Then a linearized tree was constructed for the remaining sequences according to the NJ tree topology. The divergence time (112 Mya) between B. florigae and B. belcheri was used for time calibration (Nohara et al., 2004). It should be noted that the species B. japonicum here and the amphioxus B. becheri referred by Nohara et al. (2004, 2005) are actually the same species, namely B. japonicum (Xu et al., 2005; Zhang et al., 2006).

2.5. Expression and purification of recombinant Bj-GFP1, Bj-GFP7 and Bj-GFP9

Five primers with a cohesive BamH I or EcoR I site (Table 1) were synthesized to amplify full coding sequences of Bj-GFP1, Bj-GFP7 and Bj-GFP9 (the first six amino acids of Bj-GFP1 were not amplified). PCR products were cleaved with the restriction enzymes (BamH I and EcoR I), purified, and ligated into the large fragment pGEX-4t-2 vector treated with BamH I/EcoR I enzymes. The recombinant vector was transported into BL-21 complementary cells and screened by colony PCR. After being checked by direct sequencing, the positive clones were used for preparation of glutathione S-transferase (GST)-GFP recombinant proteins. Fluorescent spectra of PBS-dissolved recombinant proteins were determined using fluorescence spectrophotometer (Cary Co., America).

3. Results and discussion

3.1. Characterization of GFP genes from three Branchiostoma species

Twenty six GFP gene models were identified from B. florigae genome database. By analyzing their neighboring genes, transcription directions and sequence identities, we can clearly distinguish alleles from different gene copies (Fig. 1). Since many gene models are not fully (without 3’ and 5’ UTR) or not correctly annotated, we re-predicted them using mVISTA and EST data. Finally thirteen functional GFP genes and two pseudogenes were determined in B. florigae (the gene reported by Deheyn et al. (2007) was named Bf-GFP9 in the present study). Among them, twelve genes (including one pseudo-gene) are distributed on two scaffolds, indicating that they are derived from frequent tandem duplication. These two clusters of genes are highly compact, just span 79 kb (scaffold_58) and 96 kb (scaffold_1) genome areas. During the preparation of this manuscript, Baumann et al. (2008) and Bomati et al. (2009) reported twelve and sixteen putatively functional GFP genes in B. florigae respectively. These controversial findings are mainly caused by the poor genome sequence of Scaffold_1 and its haplotype Scaffold_264, and also by the high sequence identities of GFP genes in this region. However our result (13 GFP genes) looks more reasonable because we consider not only the neighboring genes and sequence identities, but also gene transcription directions when we distinguish alleles from gene copies (Fig. 1). In addition, our nomenclature system is more practicable for future GFP identification from other amphioxus species, so we didn’t follow the nomenclature systems used in above two papers. To be convenient, we list all three different nomenclature systems in Supplementary Table 2. Twenty two B. lanceolatum GFP-related sequences were found from NCBI database, but at an allelic polymorphism level of 5%, they just represent six different GFP genes. Furthermore, we cloned three GFP genes from another congeneric species B. japonicum. Remarkably, all newly identified amphioxus GFP genes encode proteins of ~220 aa which is very similar to those of GFP proteins found in Cnidaria and Copepoda, suggesting that GFP protein structures are under high evolutionary constraint.

![Fig. 3. Exon–intron structures of B. florigae GFP genes. The exons are represented by boxes and the introns by stacked lines, with the length in nucleotides written above each exon and below each intron. The 5’ and 3’ untranslated regions are shown in white and the protein coding regions in black. The structures of genes marked with dots are predicted by mVISTA, not based on the alignment of cDNA and genome sequence. Question marks indicate that the size is not determined. Gene size is given on the right side of gene structure. The phylogenetic relationship between these genes is also shown. It should be noted that about 100 bp 5’ UTR sequence of Bf-GFP4 cannot be aligned to B. florigae genome sequence.](image-url)
3.2. Evolution of GFP genes in Cephalochordates

The structure of Bf-GFP9 protein has been shown to closely resemble that of GFPs in cnidarians and copepods which is characterized by a cyclic tripeptide chromophore located centrally within a conserved beta-can fold (Deheyn et al., 2007). As shown in the alignment of Supplementary Fig. 1, most amphioxus GFP proteins have the same tripeptide chromophore like that in copepod GFPs (glycine-tyrosine-glycine) but slightly different from that in cnidarian GFPs (serine-tyrosine-glycine). In contrast, the tripeptide chromophores in Bf-GFP5 and Bf-GFP6 has mutated to glycine-tyrosine-alanine. Although previous studies have shown that some mutations within the chromophore yielded functional proteins with altered fluorescent spectra (Heim et al., 1994; Delagrave et al., 1995; Mitra et al., 1996), the third amino acid glycine to alanine mutation has not been observed so far. Also, none of the seven recently examined B. floridae GFP proteins contains this mutation (Baumann et al., 2008; Bomati et al., 2009) (Supplementary Table 2). Therefore, it is interesting to determine the fluorescent spectra of these two B. floridae GFP proteins in the future.

To clarify the evolutionary relationship of the newly identified GFP genes, we reconstructed a NJ tree using their deduced protein sequences and sequences from Cnidaria and Copepoda (see Materials and methods). As shown in Fig. 2 (we did not show the original tree to save space), GFP genes from amphioxus form four major clades on the tree with each clade supported by a high bootstrap value. B. floridae GFP genes on scaffold_58 (or scaffold_408) are all included in clade 1, but those on scaffold_1 (or scaffold_264 and scaffold_549) are clustered into two clades: clade 2 and clade 4. All six B. lanceolatum GFPs fall into clade 1, but none of them has apparent orthologs from the other two species. The results suggest that 1) the above six Bj-GFPs are possibly also clustered on the same chromosome like their cognates (Bf-GFP3 to Bf-GFP7) in B. floridae, and 2) since the divergence of these two Atlantic lancelets: B. floridae and B. lanceolatum, GFP genes might have experienced a high rate of gene gain and loss. Alternatively, this phenomenon is caused by incomplete sequencing of GFP genes from B. lanceolatum. Two (Bj-GFP7 and Bj-GFP9) of the three B. japonicum GFP genes can form clearly orthologous relationship with Bf-GFP7 (in clade 1) and Bf-GFP9, Bj-ψGFP2 and Bj-GFP11 (in clade 2). We also constructed a ML tree which is similar to the NJ tree with even higher bootstraps (data not show).

The sparse and widely scattered distribution of GFP genes among animals leads to two possible explanations for their origin in
amphioxus. The first possibility is that amphioxus GFP genes have resulted from horizontal gene transfer (Bomati et al., 2009); and the other one is that the GFP family emerged before the divergence of the common diploblastic ancestor of both Cnidaria and triploblastic animals, and was secondarily lost in most examined taxa (Baumann et al., 2008). However none of them provided sound evidence to support their assumptions. The availabilities of the divergence time between B. floridae and B. japonicum (Nohara et al., 2004) and two pairs of GFP orthologous genes (GFP7s or GFP9s) from them make it possible to examine the above two hypothesis by using molecular clock method. After calibrating the clock at the nodes of GFP7 and GFP9 (Fig. 2), we calculated that the divergence time of GFP genes from Copepoda and Cephalochordata was 516 and 576 Mya respectively, which is compatible to the time (642–761 Mya) when the protostomes and deuterostomes diverged (Douzery et al., 2004). This result supports the second hypothesis, and puts GFP genes onto the most unusual evolutionary path found so far. Furthermore the clock (Fig. 2) also indicates that the four main clades of GFP genes all diverged before the radiation of the most recent common ancestor of all extant cephalochordates (162 Mya according to Nohara et al. (2005)) with the most ancient separation happening 302–340 Mya.

B. floridae GFP genes are encoded by 6–7 exons interrupted by 5–6 introns (Fig. 3). Most of them span a genome region of 3 kb long which is much shorter than the average gene length (9.1 kb) of B. floridae (Putnam et al., 2008). Their coding exons are extremely conserved although some of them have been separated for over 300 Mya (Fig. 2). This observation is consistent with their constant protein lengths.

Fig. 5. VISTA plot of the alignments between some Bf-GFP genes and their 5’ intergenic regions. Colored peaks (purple, coding; pink, non-coding and non-UTR; blue, UTR) indicate regions of at least 30 bp and 70% similarity. The phylogenetic relationship between these genes is also shown on the left. Genes from clade 4 present twice in the figure, with one showing their alignment with genes from clade 2 and the other one showing the alignment between them.
observed above re-suggesting that a high selective pressure has been acting on those genes.

3.3. Diversity of amphioxus GFP gene expression pattern

After duplication most daughter genes are gradually lost from the genome, since they are free of constraint. But some of them can survive by evolving new functions (neofunctionalization) or partitioning their ancestral functions (subfunctionalization). Obtaining a new expression profile or dividing the primary one, is an important way of neofunctionalization or subfunctionalization (Long et al., 2003; Zhang, 2003; He and Zhang, 2005). To examine whether B. floridæ GFP genes have evolved different expression patterns, we analyzed their EST data and compared their genome sequences. About 140,000 cDNA clones derived from five developmental stages of B. floridæ were sequenced from both 5′ and 3′ sides (Yu et al., 2008). Among them, 472 ESTs are related to Bf-GFP genes. After carefully clustering them with Bf-GFP cDNA (see Materials and methods), we found ESTs for eight genes (including one pseudogene, Bf-ψGFP2) (Supplementary Table 1). Actually 97.5% (460/472) ESTs are derived from two genes: Bf-GFP8 (150) and Bf-GFP9 (310), which indicates that some of Bf-GFPs have diverged their expression profile after duplication, and that Bf-GFP8 and Bf-GFP9 are both highly expressed genes when compared to average EST number per gene (262,037/21,600 = 12) in

---

**Fig. 6.** Fluorescence of two B. japonicum female individuals with different level of degenerating ovary. Pictures in the first column are taken under visible light and those in the middle and last columns are taken with a GFP filter block (Olympus, SZX-MGFP/510) and a RFP filter block (Olympus, SZX2-MRFP/BA570–620) respectively. Tissues in d–f and j–l were dissected from individual 1 (a–c) and individual 2 (g–i).
the database. Since most (376/472) ESTs are sequenced from 5' side, we discard 3'EST from further analysis. Based on 5'EST data, we find that the expression level of Bf-GFP8 and Bf-GFP9 is significantly different (Fisher-exact test, p < 0.05) in unfertilized eggs, neurula and 36 h larvae (Fig. 4a) although they are both highly expressed in these stages. Botami et al. analyzed the EST data from NCBI and reached similar result (Botami et al., 2009). The main difference is they claimed that most ESTs are from three GFP genes, corresponding to Bf-GFP8, Bf-GFP9 and Bf-ψGFP2 here. This controversial result may be caused by the exclusion of the 22 bp insertion of Bf-ψGFP2 when they clustered the GFP coding sequences with the ESTs, or by high sequence identity (98.6%) between Bf-GFP9 and Bf-ψGFP2 coding regions.

Spatio-temporal regulation of gene expression is thought to be mediated by the coordinated binding of transcription factors to discrete cis-regulatory elements. Through getting new or partially inheriting the ancestral cis-regulatory elements, duplicate genes can evolve novel or maintain primary expression profile. By comparing the genome sequences of BF-GFPs, we found that genes from different clades do not share any conserved non-coding elements (CNE, data not shown) except for those from clades 3 and 4 (Fig. 5). Interestingly, clades do not share any conserved non-coding elements (CNE, data from genome sequences of Bf-GFP evolve novel or maintain primary expression profile. Inherited the ancestral Bf-GFP9 and Bf-GFP8 (from clade 4) and Bf-GFP9 (from clade 3) have many more common CNEs (Fig. 5 and Supplementary File 1) than any other gene pair from these two clades which is consistent with the data from ESTs (Fig. 4a). Based on the high expression of Bf-GFP8 and Bf-GFP9 in embryos, we presume these elements are essential enhancers in amphioxus development. Additionally, the comparison also shows that genes in the same clade also have different CNE profiles (Fig. 5). These results indicate that after duplication BF-GFP genes have evolved divergent functions by obtaining different expression profiles.

3.4. Different fluorescent spectra of GFP proteins in B. japonicum

Using a fluorescence microscope, we observed green fluorescence in B. japonicum embryos from different development stages (data not show) which is consistent with the result reported in B. floridae (Deheyn et al., 2007). However, in adults the fluorescence is observed throughout the whole body with more concentration at the anterior end, exclusively but not always in the support cells of the oral cirri (Fig. 6b and h). This is in contrast to the hypothesis that GFP proteins are involved in photoreception or photoprotection in these sand-buried animals (Deheyn et al., 2007). In addition, we also found red fluorescence in the degenerating eggs (Fig. 6c, f, i and l). Baumann et al. also observed red fluorescence signal around oral cirri of B. floridae (Baumann et al., 2008). To further characterize fluorescent spectra of GFP proteins in B. japonicum, we expressed and purified three Bf-GFP proteins in E. coli by reconstituting them with GST proteins. When they are stimulated by UV (280 nm), both Bf-GFP1 and Bf-GFP7 have double emission peaks at 510 nm and 560 nm, and 340 nm and 562 nm respectively; and Bf-GFP9 has a single peak at 562 nm (Fig. 4b), indicating that these proteins have diversified their fluorescent spectra. Similar results were also observed in GFP proteins from B. floridae (Baumann et al., 2008; Botami et al., 2009).

4. Conclusion

The presence of GFP genes in several lineages such as Chondrichthyes, Copepoda and Cephalochordata suggests that this gene family appeared very early in the evolutionary history of the animal kingdom. But the absence of the family members in most examined deuterostomes makes the origin of GFP genes in amphioxus illusive. Our phylogenetic analysis supports that amphioxus GFP genes originated before the divergence between protostomes and deuterostomes. This indicates that GFP genes existed in the common ancestors of deuterostomes, but were lost in most examined deuterostome species, and only preserved in cephalochordates. It is one of the most unusual evolutionary processes ever found. The diversity of GFP gene family in cephalochordates can be traced back to the last common ancestor of all extant amphioxus, which had at least four GFP genes. Relatively large gene repertoire, divergent expression pattern and different fluorescent spectra make GFP in vivo functions in cephalochordates more mysterious.

Acknowledgments

The authors thank Mr. Weidong Li for his kind offer of pGEX-4T-2 vector, and Mr. Kuanchan Liu and Dr. Wei Wang for their assistance on protein expression and purification. We are also grateful to Kumars Rajali at University of Oxford for his linguistic help. This work is supported by grants from the National Natural Science Foundation of China (no. 30830023) and the National High Technology Research and Development Program of China (863 Program) [no. 2008AA092602].

Appendix A. Supplementary data


References


