European Journal of Cell Biology 92 (2013) 295–302

Contents lists available at ScienceDirect

European Journal of Cell Biology

journal homepage: www.elsevier.com/locate/ejcb

The cephalochordate *Branchiostoma* genome contains 26 intermediate filament (IF) genes: Implications for evolution of chordate IF proteins

Anton Karabinos*

SEMBID, s.r.o.-Research Center of Applied Biomedical Diagnostics, Masarykova 16, 08001 Prešov, Slovakia

ARTICLE INFO

Article history:
Received 26 July 2013
Received in revised form
17 September 2013
Accepted 10 October 2013

Keywords:
Cephalochordates
Branchiostoma
Intermediate filament
Lamin
Phylogeny

ABSTRACT

We analyzed the draft genome of the cephalochordate *Branchiostoma floridae* (*B. floridae*) for genes encoding intermediate filament (IF) proteins. From 26 identified IF genes 13 were not reported before. Four of the new IF genes belong to the previously established Branchiostoma IF group A, four to the *Branchiostoma* IF group B, one is homologous to the type II keratin E2 while the remaining four new IF sequences N1 to N4 could not be readily classified in any of the previously established *Branchiostoma* IF groups. All eleven identified A and B2-type IF genes are located on the same genomic scaffold and arose due to multiple cephalochordate-specific duplications. Another IF gene cluster, identified in the *B. floridae* genome, contains three keratins (E1, Y1, D1), two keratin-like IF genes (C2, X1), one new IF gene (N1) and one IF unrelated gene, but does not show any similarities to the well defined vertebrate type I or type II keratin gene clusters. In addition, some type III sequence features were documented in the new IF protein N2, which, however, seems to share a common ancestry with the *Branchiostoma* keratins D1 and two keratin-related genes C. Thus, a few type I and type II keratin genes existed in a common ancestor of cephalochordates and vertebrates, which after separation of these two lineages gave rise to the known complexities of the vertebrate cytoplasmic type I–IV IF proteins, as well as to the multiple keratin and related IF genes in cephalochordates, due to multiple gene duplications, deletions and sequence divergences.

Introduction

The filamentous IF network in metazoan cells seems to be responsible for resistance against mechanical stress (*i.e.* Mclean and Lane, 1995; Hesse et al., 2000; Karabinos et al., 2001b; Vijayaraj et al., 2009; Zhang et al., 2011). There are about 70 different members of the IF protein family in man (Hesse et al., 2000, 2004) and in vertebrates, which are subdivided into the five major types (for reviews see Fuchs and Weber, 1994; Parry and Steinert, 1995; Herrmann et al., 2003, 2009). Keratin epithelial filaments are based on obligatory heteropolymeric double-stranded coiled coils, each dimer containing one type I and one type II chain. The four mesenchymally expressed type III proteins – desmin, vimentin, GFAP and peripherin – generally form homopolymeric IFs. The Type IV are three neurofilament proteins and α-internexin, while in the type V group are the nuclear lamins. The two eye lens IF proteins – filensin and phakinin – fall outside these types.

All IF proteins possesses a central rod domain containing heptad repeats which is flanked by variable head and tail domains. The structure of the rod domain in IF molecules enables them to assemble both *in vitro* and *in vivo* into one of four closely related 10 nm-like filaments (Fuchs and Weber, 1994; Parry and Steinert, 1995; Herrmann et al., 2003, 2009). The central rod domain of all IF proteins is subdivided into segments 1A, 1B, 2A and 2B, however, the nuclear lamins and the protostomic cytoplasmic IFs contain a longer rod segment 1B. In addition, the nuclear lamins have a unique tail containing an Ig-like segment, a nuclear localization signal and, in most cases, a CaaX box (Erber et al., 1999). It is assumed that lamins represent an ancestor sequence of cytoplasmic IFs (Fuchs and Weber, 1994; Parry and Steinert, 1995; Erber et al., 1998; Herrmann et al., 2003).

In our previous studies on cephalochordate *Branchiostoma* we characterized 13 cytoplasmic IF proteins. Five proteins were identified as *bona fide* keratins forming the obligatory heteropolymeric IF from mixtures of any type I (k1, Y1, E1) and type II (D1, E2) recombinant proteins. In addition, two of the *Branchiostoma* type I keratins polymerize also with the human type II keratin 8 (Karabinos et al., 2000). Three keratins (k1, Y1, D1) and protein X1 are expressed in the gastrula. The number of lancelet IF proteins increases at the neurula and early larval stages to 7 and 11 respectively, and in the adult 13 different proteins have been found. The keratins are the major IF proteins in the *Branchiostoma* nerve cord. Proteins X1, C1 and C2 possess some keratin-like characters and were shown to be integrated into the epidermal and neuronal keratin meshwork

* Tel.: +421 905 402 683.
E-mail address: sembid.pp@gmail.com

0171-9335/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ejcb.2013.10.004

(Karabinos et al., 2001a). Finally, the five remaining *Branchiostoma* IF proteins A1, A2, A3, B1 and B2 formed a separated A/B branch in the evolutionary trees and were proposed to be lancelet-specific (Karabinos et al., 2002). The B1 protein is expressed in mesodermally derived muscle tails and in coelomic epithelia and forms homopolymeric IF *in vitro*. In contrast, its closest relative B2 is co-expressed with the three homologous proteins A1–A3 in the intestinal epithelia and can form heteropolymeric IF with A3, driven by a putative trigger-like sequence in segment 1B (Karabinos et al., 2002, 2012).

Thus, the only homologs of the vertebrate cytoplasmic type I to II IF proteins, identified so far in cephalochordates, contrast with the situation in urochordates, a sister group of vertebrates (Putnam et al., 2008 and references therein), where also a vertebrate type III homologous IF protein was found (Wang et al., 2000; Karabinos et al., 2004). It is therefore possible that additional *Branchiostoma* IF proteins might have escaped previous cDNA cloning and that also this early-diverging chordate contains some vertebrate type III and, eventually, also type IV homologs. Moreover, a comprehensive analysis of the cephalochordate IF complement can also be used to test the recent conclusion that all type I and type II tetrapod keratins evolved from only two genes that were present in the ancestor of extant vertebrates (Vandebergh and Bossuyt, 2012).

In this study we identified 13 newly predicted IF sequences in the *B. floridae* draft genome (Putnam et al., 2008). None of them was defined as a vertebrate type III or type IV homolog. These data indicate the existence of a multigene family of IF proteins in the cephalochordate *Branchiostoma* which evolved independently from the multiple type I–IV IF proteins in vertebrates.

Results and discussion

Identification of 26 IF genes in Branchiostoma genome

We made BLAST searches of the *B. floridae* predicted proteome (Putnam et al., 2008) using the 13 previously cloned *Branchiostoma* cytoplasmic IF sequences A1-A3, B1, B2, C1, C2, D1, E1, E2, k1, Y1, X1 (for references see Introduction) and the lamin (Riemer et al., 2000) as a query. In total, 26 predicted IF-like proteins or fragments were identified. Comparison of all these sequences with the corresponding genes in the *B. floridae* genome (Putnam et al., 2008) found one prediction (BRAFLDRAFT_123713) which covered two neighboring genes and two other predictions (BRAFLDRAFT_132259 and BRAFLDRAFT_116897), representing allelic variants of one gene. We used manual corrections in several protein predictions based on their comparisons to previously reported *Branchiostoma* IF genes and protein sequences. However, identification of final amino acid sequences of the terminal head and/or tail domains for four genes (224304, 123713-E2, 123713-E3 and 235856; Table 1) awaits cloning of their corresponding full length cDNAs. 26 IF genes, identified in *Branchiostoma*, contrast with six such genes found in the draft genome of the urochordate *Ciona* (Karabinos et al., 2004), which, however, is thought to have undergone a substantial gene loss (Putnam et al., 2008). Table 1 summarizes our search and shows that all previously identified IF proteins (see Introduction), except A2, were identified in the *B. floridae* genome. Pairwise analyses of the amino acid rod sequences of the newly identified IF sequences and the corresponding previously reported *B. floridae* IF proteins (denoted with an "Bf" prefix in Table 1) revealed divergence ranging from 1% (E2) to 4% (A1). Moreover, as mentioned above, no obvious counterpart of the previously reported *B. floridae* A2 IF protein could be identified in the *B. floridae* draft genome (Fig. 1A and text below). These results support high allelic variations of the *B. floridae* genome (Putnam et al., 2008) and might also indicate an existence of different *B. floridae* sub-populations, as recently documented for the related pacific lancelet *B. belcheri* (Li et al., 2013). However, due to the current ambiguities regarding the *B. floridae* IF protein A2, we did not apply its name for the nomenclature of the A-type IF sequences, identified in the *B. floridae* draft genome. Finally, from the remaining 13 predicted IF sequences not reported before, four belong to the previously established *Branchiostoma* IF group A (A4-A6 and AΨ), four to the *Branchiostoma* IF group B (B2b–B2e), one is homologous to the type II keratin E2 (E3), while the analyses of the last four new IF sequences N1 to N4, which could not be readily classified in any of the previously established *Branchiostoma* IF groups, are described below (Table 1, Fig. 3). Notably, the EST analysis document transcription activity for only five (A6, B2e, E3, N1, N2) of the 13 new IF sequences (Table 1), which left open the question whether the remaining eight proteins are also active genes. We propose here, that at least the gene AΨ does not encode a functional IF protein (see below), but a more precise picture about the functionality of this and other newly identified IF genes will be established once various gaps of different lengths in these sequences are closed (see Table 1 for details).

All keratin-like A and B2-type IF genes are clustered together

As documented in Table 1, 12 (A1, A3-A6, AΨ, B1, sB2a-B2e) of 26 IF sequences identified in the *B. floridae* draft genome, are related to the two previously established *Branchiostoma* keratin-like A and B-type IF groups. Pairwise amino acid and distance-based phylogenetic analyses, using either the unweighted pair-group method with arithmetic mean (UPGMA; Fig. 1A) or the neighbor-joining method (data not shown), of the rod domains of the new and previously reported A/B-type sequences (Riemer et al., 1998; Karabinos et al., 2002) indicate that the predicted sequences 81360 (A1) and 265949 (A3) correspond to the previously reported *B. flordae* IF proteins A1 and A3, respectively, and that the A4 (81363), A5 (81358) and A6 (224304) sequences are their homologs. The latter is true also for the predicted sequence 81362, which, however, is a partial sequence due to an stop codon positioned in the region encoding the rod segment 2B as well as due to deletion in a distal part of the corresponding gene (Table 1). We think, therefore, that the 81362 sequence represents a non-functional A-type gene (*i.e.* pseudogene), which we therefore named here AΨ (Table 1, Fig. 1A). As mentioned above, none of the sequences analyzed in Fig. 1A is an obvious counterpart of the previously reported *B. floridae* A2 (Fig. 1A). The phylogenetic tree in Fig. 1A further reveals that the B2a (81357) sequence corresponds to the previously cloned *B. floridae* IF protein B2 and that the sequences B2b (81363), B2c (224428), B2d (224321) and B2e (281325) are its homologs. Finally, the predicted sequence 77526 corresponds to the previously reported IF protein B1 (Table 1), which is currently the only *Branchiostoma* IF proteins able to form homopolymeric IFs (see Introduction for details). Interestingly, this and all the other identified A/B-type sequences here possess a trigger-like motif in segment 1B which has the potential to form multiple salt bridges (Fig. 1B) and probably also to trigger formation of obligatory heterodimer as previously demonstrated for the A3 and B2 polypeptides (Karabinos et al., 2012).

Fig. 1C documents that all predicted A and B2-type IF genes described above, are grouped together on the *B. floridae* scaffold 101 while only B1 is localized on the separated scaffold 71 (Table 1). The A/B gene cluster on the scaffold 101 still shows a variety of gaps but it can be clearly seen that the three B2-type IF genes B2e, B2c and B2a are individually paired with the three A-type IF genes A3, A6 and A5, respectively. Moreover, the five genes B2e, A3, B2c, A6 and A5 from the latter six have their evolutionary most closely related counterparts (see tree in Fig. 1A) B2d, A1, B2b, AΨ and A4, respectively, on the opposite side of the scaffold (indicated in Fig. 2C by brackets under sequences), while only the B2a gene lacks such a