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

counterpart on the opposite side of the scaffold (marked by "?" in Fig. 1C).

Based on the data described above and the same gene structure of all A/B-type IF proteins presented here (Fig. 3B; Karabinos et al., 2000), we have constructed a hypothetical scheme showing their origin due to repeated duplications in the *B. floridae* genome. The single archetypal ancestor of B and A genes was duplicated into B1/B2 and A genes. The former gene was subsequently duplicated to the B1 and B2 genes which was followed by a B1 genomic transposition from the original place in the genome (Table 1). The two ancestral genes B2 and A, which were left (on the scaffold 101), were subjected to two rounds of tandem gene duplications, leading to three ancestral B2/A gene pairs. This was followed by a duplication of the whole six-gene cluster, deletion of a hypothetical B2a gene counterpart with an attached part of the neighboring AΨ gene (Table 1) and by a sequence divergence leading to the origin of 11 A/B-type IF genes in the *B. floridae* genome (Fig. 1D).

Clustering of keratins and related IF genes in Branchiostoma genome

As mentioned in the Introduction, the vertebrate type I to II keratin homologs were previously identified in the cephalochordate *Branchiostoma* (Karabinos et al., 1998, 2000). Here we found that genes for the two type I keratins E1 and Y1, the type II keratin D1, the two keratin-like proteins X1 and C2, as well as the new IF N1, are clustered together on the genomic scaffold 328 (Table 1, Fig. 2A). This *Branchiostoma* keratin-like cluster is interrupted by one predicted IF unrelated gene (98782) and does not show any similarities to the well-defined vertebrate type I or type II keratin gene clusters (Zimek et al., 2003; Zimek and Weber, 2005; Hesse et al., 2004; Vandebergh and Bossuyt, 2012). Thus, the typical organization of keratins and keratin-flanking genes of vertebrates originated after divergence of *Branchiostoma* and vertebrates. The remaining previously reported *Branchiostoma* type I keratin k1 is located on the scaffold 664 (Table 1), while the *Branchiostoma* type II keratin E2 is attached on the scaffold 126 in tail-to-head orientation to the new IF gene which we named here E3 (Table 1). Both latter genes are only about 5,5 kb apart from each other, have identical gene structure (Fig. 3B) and the 73% sequence identity (Table 1), which collectively suggest their origin by a relatively recent gene duplication. Thus, these data and the phylogenetic (Fig. 4) and EST analyses (Table 1), suggest that the E3 IF protein represents, beside D1 and E2, the third type II keratin expressed in the cephalochordate *Branchiostoma.*

Branchiostoma IF proteins C, D1 and N2 have a common ancestry

Previously we reported that the protein C2 from *B. lanceolatum* forms with keratins the epidermal and neuronal IF meshwork (Karabinos et al., 2001a) and also possesses some keratin-like sequence features (Karabinos et al., 2002). Moreover, both C2 and its closely related *B. floridae* IF protein C1, contain a long unique

Table 1
The complement of IF proteins of the cephalochordate *B. floridae*.

JGI hypothetical protein BRAFLDRAFT_	Scaffold	GenBank acc. no. EEN_	Branchio-stoma IF protein	Corrections of JGI hypothetical protein prediction and (note)	Best rod AA sequence identity of*	cDNA or EST data*
Lamin						
121709	71	42741.1	Lamin	EXT Exon 7 (22 AA)	95% to BlLam	cDNA
A-type IF sequences						
81360	101	57330.1	A1	None	96% to BfA1	cDNA
265949	101	57324.1	A3	None	98% to BfA3	cDNA
81363	101	57333.1	A4 (new)	EXT Exon 2 (14 AA)	91% to A1	–
81358	101	57328.1	A5 (new)	None	88% to A4	–
224304	101	57326.1	A6 (new)	None (Gaps between Exons 3 and 4)	Exons 1–3 96% to AΨ, Exons 4–6 96% to A1	–
81362	101	57332.1	AΨ (new)	None (only Exons 1–5)	96% to A6	–
B-type IF sequences						
77526	71	50569.1	B1	TRM Exon 1 (45 AA) TRM Exon 6 (46 AA)	100% to BfB1	cDNA
81357	101	57327.1	B2a	TRM Exon 4 (7 AA) (partial: Exons 1–4; Gaps past Exon 4)	97% to BfB2	cDNA
81361	101	57331.1	B2b (new)	None	94% to BfB2	–
224428	101	57325.1	B2c (new)	EXT Head (28 AA)	94% to BfB2	–
224321	101	57329.1	B2d (new)	EXT Head (26 AA)	94% to BfB2	–
281325	101	57323.1	B2e (new)	None	94% to BfB2	EST
Keratin and related IF sequences						
129989	403	43152.1	C1	None	100% to BfC1	cDNA
235821	328	45673.1	C2	EXT Head (117 AA) EXT Tail (47 AA)	98% to BfC2	cDNA
61222	328	45675.1	D1	None	89% to BfD1	cDNA
160578	328	45671.1	E1	Added Exons 1–4,7	98% to BfE1	cDNA
123713	126	62249.1	E2	TRM C-term (544 AA)	99% to BfE2	cDNA
123713	126	62249.1	E3 (new)	TRM N-term (510 AA)	73% to BfE2	EST
132259	664	42329.1	k1	None	98% to Bfk1	cDNA
235856	328	45672.1	Y1	EXT Head (67 AA)	90% to BlY1	cDNA
117261	3	45676.1	X1	None	86% to BlX1	cDNA
Other IF sequences						
128770	328	45670.1	N1 (new)	None	28% to BlY1	EST
124339	145	57617.1	N2 (new)	None	38% to BfB1	EST
88547	175	49529.1	N3 (new)	None	22% to BfB1	–
74516	51	54819.1	N4 (new)	None (gaps between Exons 5–6)	28% to BFD1**	–

Genes were identified by our screen of the *B. floridae* draft genome (Putnam et al., 2008).

* Previously reported *B. floridae* and *B. lanceolatum* IF proteins, denoted with the prefix "Bf" and "Bl", respectively, and the corresponding cDNAs are from Riemer et al. (1998, 2000), and Karabinos et al. (2000, 2001a).

** This most divergent *B. floridae* IF sequence displays 28% sequence identity in the coil 2 segment while in the coil 1 sequence the identity value to various *B. floridae* IF sequences is below 14%; EXT, extended; TRM, trimmed; AA, amino acid; N-term, N-terminus; C-term, C-terminus.

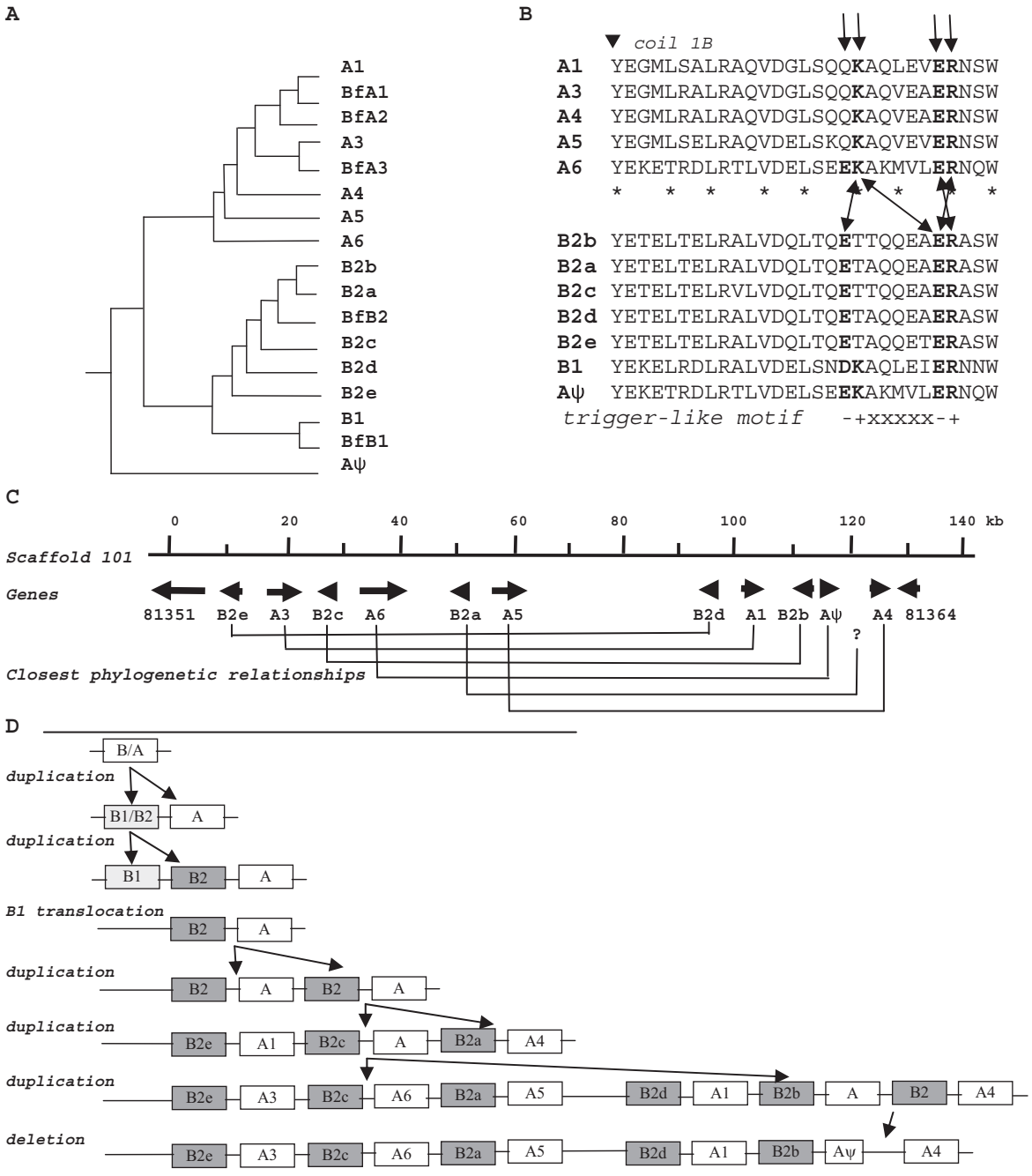

Fig. 1. (A) An evolutionary tree generated by the analysis of the conserved rod domains of various *B. floridae* A and B-type IF sequences. Sequences denoted by a "Bf" prefix were reported previously (Riemer et al., 1998), other sequences were identified here in the *B. floridae* draft genome. This unrooted tree is constructed by UPGMA analysis (see "Materials and methods" for details). Note that the sequence corresponding to the previously reported protein A2 (BfA2) has not been identified in the *B. floridae* draft genome (see text for details). (B) Sequence alignment of the first 29 residues of segment 1B of various *B. floridae* A and B-type IF proteins identified in this study. Arrows pointing down mark the start of segment 1B and asterisks mark the *a* and *d* positions of the heptad repeat pattern. The four arrows above the sequences mark the putative trigger-like motif "−+xxxxx−+" identified previously in the heteromeric IF proteins B2 and A3 (see Karabinos et al., 2012 for details). Residues of the aligned sequences that fit the charged positions of the "consensus trigger-like motif" are indicated in bold. (C) Schematic representation of the A- and B2-type gene cluster in the *B. floridae* genome. The genes are referred according to their location on the scaffold 101. The presented IF cluster is flanked by the IF unrelated genes 81351 and 81364. The gene pairs revealing the closest sequence relationships in the phylogenetic tree, presented in panel C of this figure, are connected by brackets below. The region on the scaffold where a counterpart of B2a is lacking is indicate by? (See text for details). (D) Proposed scheme for the evolution of the six A-type and the five B-type IF genes in the *B. floridae* genome (see text for details).

tail with two degenerate repeats exhibiting a heptad repeat pattern compatible with a coiled coil formation ability (Riemer et al., 1998). Interestingly, we found that the C2 counterpart, identified here in the *B. floridae* draft genome, lacks a long tail domain (Fig. 2B) and resides on the *B. floridae* genomic scaffold 328 with genes for the keratins E1, Y1, D2, the keratin-like IF protein X1, the new IF N1 as well as the IF unrelated gene 98782 (Fig. 2A). However, analysis of the latter gene 98782, located on the scaffold 328 downstream to C2 (Fig. 2A), reveals a strong sequence similarity to the tail sequences of the *B. lanceolatum* C2 and the *B. floridae* C1, D1 and N2 IF proteins (Fig. 2B). Moreover, all latter sequences, except C2, end with the "ricin-type" carbohydrate-binding domain which is found in a variety of molecules serving diverse functions such as enzymatic activity, inhibitory toxicity and signal transduction

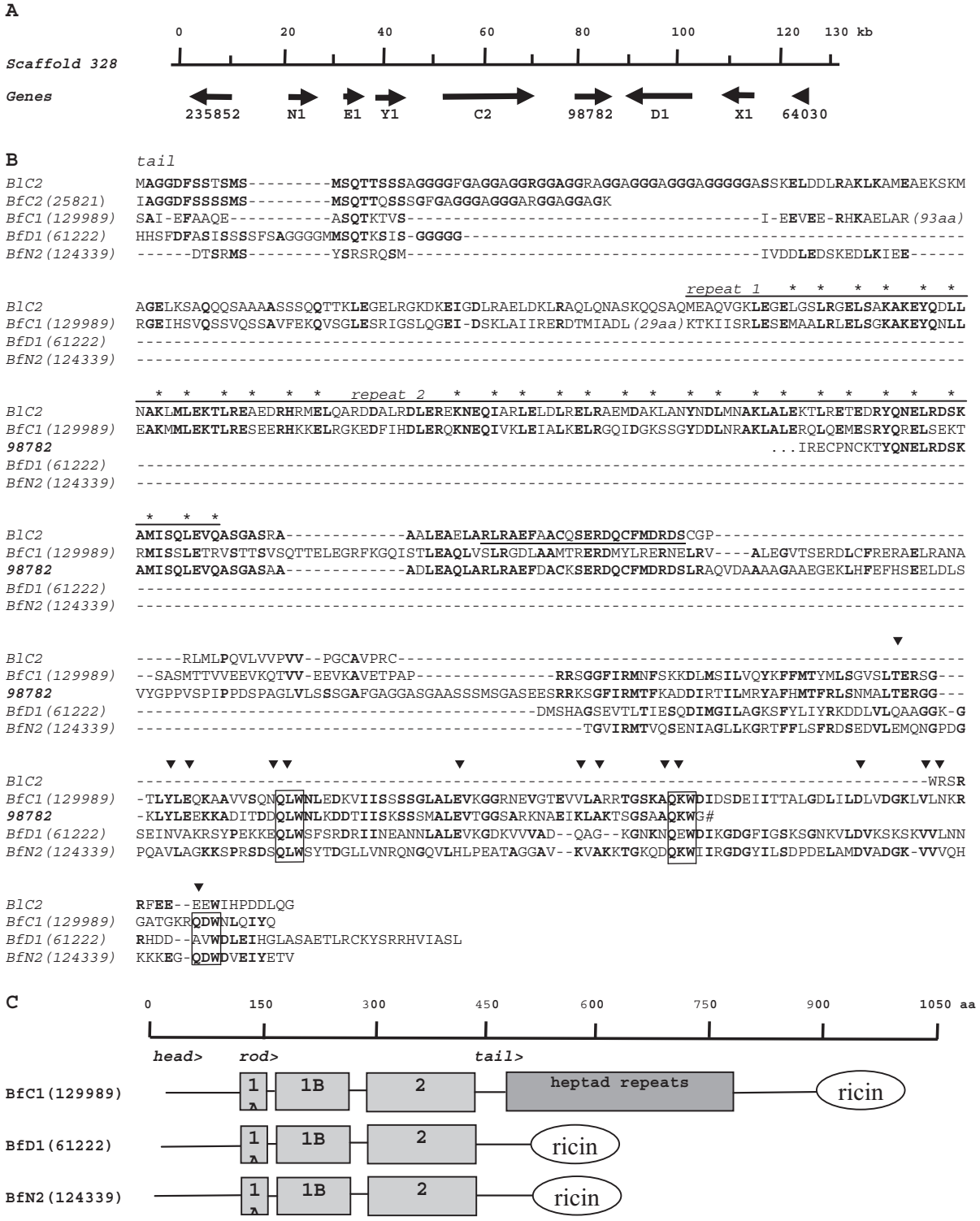

Fig. 2. (A) Schematic representation of the gene cluster in the genome of *B. floridae* which contains genes for various keratins, keratin-like IF proteins and one IF unrelated gene (98782). The genes are referred according to their location on the scaffold 328. The presented IF cluster is flanked by the IF unrelated genes 235852 and 64030. Note that the third IF unrelated gene 98782 of this cluster reveals strong sequence similarities to the tail domains of the *B. floridae* D1, C1, N2 and of the *B. lanceolatum* C2 IF proteins (see panel B and text for details). (B) Alignment of the amino acid sequences of the entire tail domains of the *B. floridae* (C1, C2, D1, N2) and *B. lanceolatum* (C2) IF proteins and the carboxyterminal portion of the predicted IF unrelated *B. floridae* protein 98782. Identical amino acids are marked in bold, while dashes are used to optimize the sequence alignment. Asterisks indicate the primarily hydrophobic residues present in the *a* and *d* positions of consecutive heptads of the C1 and C2 tail domains (Riemer et al., 1998). The "QxW" motifs of the "ricin-type" carbohydrate-binding domain of the *B. floridae* proteins C1, D1, N2 and 98782 are boxed, while positions of the 14 putative sugar binding sites presented in the canonical "ricin-type" carbohydrate-binding domain are marked by arrowheads pointing down (see text for details). Note that in the *B. lanceolatum* C2 sequence only a few last residues reveal some similarities to the "ricin-type" domain. Note also that we corrected here an open reading frame of the original C2 tail sequence (Riemer et al., 1998) by exchanging the 22 residues "ACVPSLLPASPSATSASWTATR" for the 22 residues long sequence "RLRAEFAACQSERDQCFMDRDS" which we previously overlooked due to the frame-shift error (our unpublished data). This new sequence part of the *B. lanceolatum* C2 is underlined. (C) Schematic representation of the three unique *B. floridae* IF proteins C1, D1 and N2. The structural organization of an IF protein consists of a head, a rod and a tail domain, as indicated. The helical rod covers segments 1A, 1B and 2. The tail "heptad repeats" domain of C1 as well as the "ricin-type" carbohydrate-binding domain of all three presented proteins have not been observed in a large collection of the currently known IF proteins (see text for details).

A

```
head
N2                                                                                          MSRSGTQSTTTLSHTVPVQRPGPRPSATS
N3   ...YGLKRSPDREFKVFKREAYGSYYRSLQAGYSGRQDDTDSQASVESWKGMRIGPPKKPRVSSTSSERGFDMRRDRESYRGHDTSFGGYSRERRPSGDSVQSVTEGM

N1                                                                         MSFMTGGFGGGGGGGGGGGGGGGAGLSLTKSGGGGGGGMNSRQEL
N2   TPTKTRLSGSPRSMTDSIYAARFSTSPAMSNYNSATKTLTTTYRLGGGGTNNRTSFSGSRPSPVTTMRRTTIAASRPSTTTISGAIATSNALQIAENHALRVQEKEEM
N3   LERRRFDRDRGPKSPLDEAWDRDRAQGYRDRRFSEDRGYRDRDRGTGTYRDRGYSESDRKDLGRGLDLSSDRHRDLEPERDRDHYRDREHDRDRDRDYKRSSRDLSRL
N4                                                                                          MGTDHLSLWVENVVCYCDQNISIALQSLP

rod ▼  coil 1A                          ▲             L1        ▼  coil 1B
N1   NTLNTRLSGYINKVRSLEARNAELEMLIAEASKAHAAAGVSTAPISIDTSELMAMKKEIQATANEVSYLESQKSTLLIEIENLRRQLEDLLRRRAEMESDISGAKGEL
N2   QTLNQRLAGYIDKVRFLETRNEELEIEVRALQAGHQNFHQMKE---MFEQEIRELRRAIDSLSNEKAIVEVDRDTLGEELDHTRHQLGEEKRVRQDLETELNSSRKEL
N3   EEGGEEAKRYIEKIQELEDRQDYLEDKIRALGGRPEPLSRAEE---DLQQEVEEAHTELDAITVLKAKLETEVNNLRKDVEEHRRRVDGERQARYAAEETFRKVQAEY
N4   YKEIRQGDLYFPGHLDTPSITFDTRRTCARATTVNVPPDLWPR---IRPTDVNGTRICGRVYEEGAPVLGGERLLPPPREQAPAHVFALRQTALFWSARKDEHVVMSS
      *   *  *   *  *   *  *   *  *                *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *

                                             ▲               L12              ▼ coil 2A   ▲   L2    ▼ coil 2B
N1   DSALGALDGLQDEVAQAEDEMSSAAQGPAQATTIEE-------------QIDVVQVE---RLRADLNDSLNALRGEFEQLASQGIADVEGDWEARLRELQAEIERLKA
N2   NKVDLARQDMETRLKTALQDLDFSKEIHQQEIEALQKQLE---------EQKAVVQIDSGRSIYDEMAMAMANMKKQYEALAKKTKDESEQWYQSKLRDITEKCDRAED
N3   SELRGIIDDINGQIDASRERLRAIKYEHNEEKRHLQSRL----------GDDTGSDRSGGS-YTHYSRDLYNYRRRDTHSREHHDYDSSA-NKRRLFEMMASVEMEVN
N4   IAWAVSGPGWQARTPSARALFPTPPTLRPALPSIKDKQAEAALKKKMEELSSCVQ-----KYHVELTETLQRMSVQQEDSARMSKQELEQWYRTQLEEAKAAAAKDAE
      *   *  *   *  *   *  *   *  *   *                                      *   *  *   *  *          *   *  *   *  *

                                           stutter                                                                ▲
N1   LLEELRMQISQLRAELSAILSEAEQIKATNISLERDISTLKDTIRKRLEDYRIQIADLNGEGQRLDGEIQAVVAQ----YNDLLNKKLALEKEVIDYEQIVGGENQRL
N2   ALRLARQEALENKRQVHVLTSEVDTAKSKNALLEKSARELESRSTGDVDQLQSMVRSLEDELKQIKNEMAKQVQD----YNELMNVKLALDVEIQAYRKLLEGEETRM
N3   ELQRVRREIVRQQELVRTLHIDIHTLQFREDELRKVANDTDKRHGREVDIRDNVVTKLQQEVASATQELLQQKYQ----NQQLLDDAMLLDAQLKTYRRLWDLSEGRG
N4   TLTETRDALLQSETLVGHLSDEIDMLRSQNSSLSKQLQEAEKAVDENTVKYEENIRVLETELQALKEDTATQIMEIVHHVMSLTNARVALDLQMDSYFDILGKRKSDT
     *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *  *   *
tail
N1   KVQGGGYAIQQISIGSGSSSGGLGGGMGFGGGGGGFGGGGGFGGGGGFGAGGGGGFGGGGGGGFSKGGGGGGGFSRGGGGSRGGGFSKGGGIKFNL
N2   DTSRMSYSRSRQSMIVDDLEDSKEDLKIEETGVIRMTVQSENIAGLLKGRTFFLSFRDSEDVLEMQNGPDGPQAVLAGKKSPRSDSQLWSYTDGLLVNRQNGQVL...
N3   PGDPSSAVGDEEETT
N4   TEADTVKKQKSLVLPALPIAEDNRVPMDTSDACAQDCLSDENANTSPIMADEVAAGSGH
```

B

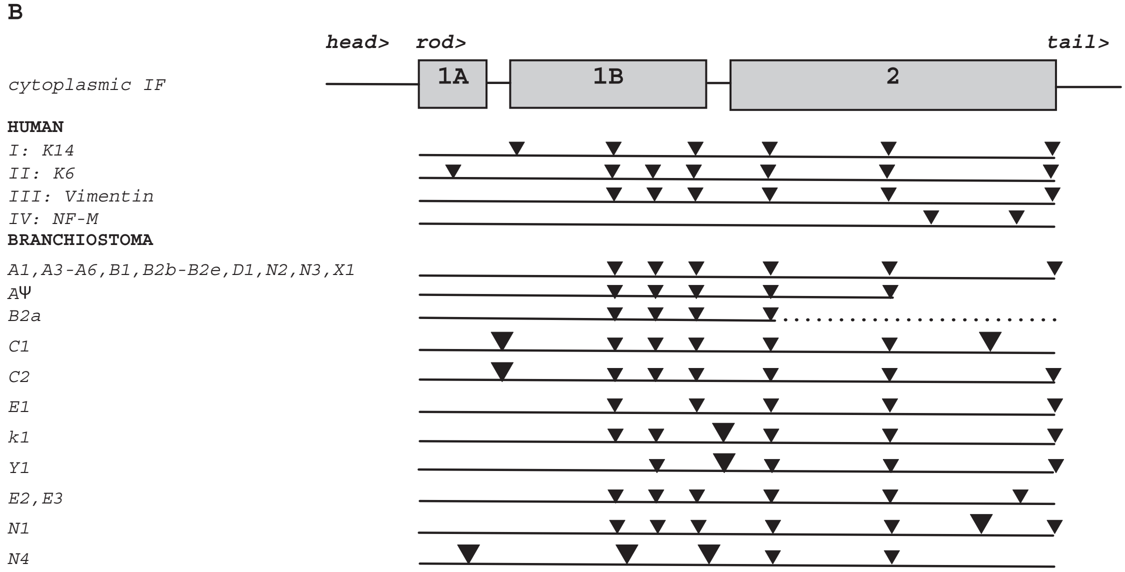

Fig. 3. (A) Alignment of the amino acid sequences of the new *B. floridae* IF proteins N1, N2, N3 and N4 identified in this study. The head, coil 1 and the linker L12 are indicated. Arrows pointing downwards and upwards mark the beginning and the end of the rod segments 1A, 1B, 2A and 2B which are connected by linkers L1, L12 and L2. Asterisks below the rod segments mark the *a* and *d* positions of the heptad repeat pattern. Dashes are used to optimize the sequence alignment. Bold letters indicate identical residues in at least two aligned rod sequences. Note that the N4 rod coil 1 lack the usual number of apolar residues in the *a* and *d* positions of heptads and contains also a number of prolines. Moreover, the coil 2 segment of N4 contains a unique 4 residue long insertion. Also note that both the terminal head and tail domains of N1 contain, like those of several cephalochordate and mammalian keratins, several glycine loops flanked by hydrophobic residues (indicated in bold letters). The beginning of the unique "ricin-type" carbohydrate-binding domain in the N2 tail segment is underlined. (B) Comparison of intron positions of the rod domain in 25 cytoplasmic IF genes of *B. floridae* and of vertebrate IF type I to IV genes. The tripartite structural organization of the rod domain of the cytoplasmic IF proteins is indicated at the top. Dots indicate part of the AΨ and B2a genes not yet analyzed. Small and large filled arrowheads specify common and unique intron positions, respectively. Gene structures of the corresponding *B. lanceolatum* IF genes were referred elsewhere (Karabinos et al., 2000).

and which contains three "QxW" motifs (boxed in Fig. 2B) and 14 putative sugar binding sites (marked by arrowheads pointing down in Fig. 2B; for review see Hazes, 1996). Based on this definition is the latter domain complete in the C1 and N2 proteins but partial in the protein D1 and the IF unrelated protein 98782. A described sequence conservation pattern in tails, which we previously did not observe (Riemer et al., 1998; Karabinos et al., 2000), indicates a common ancestry of all these IF proteins C1, C2, D1 and N2 from a keratin-like ancestor containing a "ricin-type" carbohydrate-binding domain. This ancestor, then, due to gene duplications and a subsequent domain deletion and sequence divergence, gave rise to the type II keratin D1, the two keratin-like proteins C1 and C2, the new *Branchiostoma* IF protein N2, as well as to the IF unrelated protein 98782, identified now in the *B. floridae* draft genome. Thus, a unique globular structure of the latter four IF proteins (Fig. 2C) was not observed in a large collection of currently

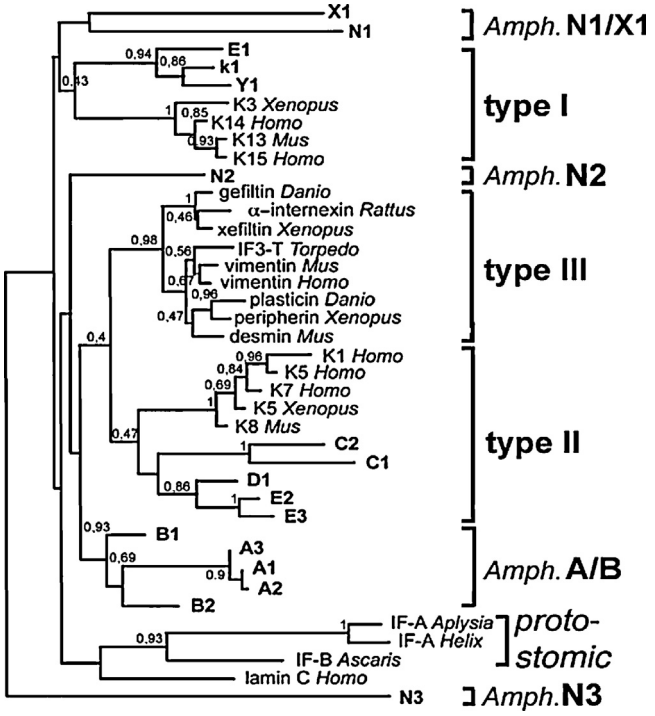

Fig. 4. Phylogenetic analysis of *Branchiostoma* and other IF rod sequences from invertebrates and vertebrates using the maximum likelihood method. The bootstrap values (100 replications) are shown above the internal nodes. The four subfamilies I–IV of vertebrate IF proteins as well as the position of the *Branchiostoma* A/B, N1/X1, N2, N3 and the protostomic IF sequences are indicated.

known protostomic and vertebrate cytoplasmic IFs. Moreover, the data regarding the *B. lanceolatum* long tail C2 protein *versus* its short-tail counterpart in *B. floridae* also document that a diversification of the multigene family of IF proteins continues in both *Branchiostoma* lineages.

Analysis of four new Branchiostoma IF protein sequences N1 to N4

All four yet unclassified IF proteins N1 to N4, found in the *B. floridae* draft genome (Table 1, Fig. 2 and text above), contain a typical IF-like central rod domain with segments 1A, 1B, 2A and 2B characterized by α-helices with heptad repeats which favor coiled-coil formation (Fig. 3A). An exception is the N3 rod starting and ending with the poorly conserved helix initiation and termination motif (Fuchs and Weber, 1994; Parry and Steinert, 1995), respectively, and the same holds also for the N4 rod sequence. Moreover, the whole coil 1 segment of N4 exhibits very poor heptads due to many polar residues at the heptad positions *a* and *d* and 14 prolines, while its coil 2 segment reveals somehow unusual four-residue insertion (Fig. 3A). Interestingly, also the corresponding N4 gene exhibits a unique exon-intron feature in contrast to the high conservation of other *Branchiostoma* and vertebrate IF genes (Fig. 3B). However, one unique intron could also be seen in the gene N1 (Fig. 3B; see also Karabinos et al., 2000). The rod domain of the N1, N2, N3 and N4 proteins is N-terminally flanked by a variable head of 45, 137, 363 and 29 residues, respectively, and C-terminally by a tail of 97, 173, 15 and 59 residues, respectively (Fig. 2B and 3A). Interestingly, both terminal domains of the predicted N1 protein contain, like those of several cephalochordate and mammalian keratins, several glycine loops flanked by hydrophobic residues (see Fig. 3B). On the other hand, the predicted N2-tail, as mentioned above, shares similarities with those of the D1 and C1 proteins, including a presence of the "ricin-type" carbohydrate-binding domain (see Figs. 2 and 3A). Finally, the head and tail domains of the predicted N3 and N4 proteins have no significant similarity to any other protein in the NCBI non-redundant protein database.

Fig. 4 shows an evolutionary tree generated from the sequences of the rod domains of the four new *Branchiostoma* IF proteins N1 to N4 (see above) and a variety of representatives of *Branchiostoma*, protostomic and vertebrate type I-IV IF proteins (Table 1; Karabinos et al., 2002). The N1 sequence relates in this tree to the keratin-like *Branchiostoma* protein X1 (Karabinos et al., 2001a, 2002) and to the group consisting of the *Branchiostoma* and vertebrate type I keratins. N1 might therefore represent a new keratin-like IF in *Branchiostoma* in line with its keratin-like sequence features of the terminal domains (Fig. 3A and text above) and its genomic clustering with keratin genes (see Fig. 2A and text above). The N2 protein, on the other hand, displays in the phylogenetic tree a sister group relationship with the type II, type III and the *Branchiostoma* A/B groups (Fig. 4). This evolutionary relationship and a sequence conservation pattern in an alignment of its tail domain with those of D1 and C IF proteins (see Fig. 2 and text above) support a common ancestry of N2 with these three keratin and keratin-like IF sequences (Karabinos et al., 2002). However, N2 seems to have gained, independently from the type III proteins in vertebrates, also some vimentin-like characteristics as revealed by a BLAST search of this sequence against the NCBI non-redundant protein database, where it shows similarities exclusively to vimentins (data not shown). Finally, the predicted protein N3 (Fig. 4) exhibits the most remote IF sequence among the proteins used in constructing the phylogenetic tree (Fig. 4) and the same holds also for the fourth new IF protein N4 which, due to its aberrant sequence (Fig. 3A), was not included in the present phylogenetic analyses. Neighbor-joining distance analysis of the same sequences resulted in a tree similar to that in Fig. 4, however, none of the N1 to N3 phylogenetic position had bootstrap support in any of the phylogenetic analyses. Thus, future experiments, including expression and assembly studies are expected to improve the classification of the four new IF proteins N1 to N4 and to provide clues to their functional significance in *Branchiostoma*.

Taken together, the above presented data collectively indicate that a few type I and type II keratin genes existed in a common ancestor of cephalochordates and vertebrates, which after separation of the two lineages gave rise, due to multiple gene duplications, deletions and sequence divergences, to known complexities of the cytoplasmic IF proteins in vertebrates (Hesse et al., 2004) as well as to the multiple keratin and keratin-like IF genes in the cephalochordate *Branchiostoma* (Karabinos et al., 2000, 2002 and this study). This finding supports and extends 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). Thus, the vertebrate type III proteins seems to have appeared first in a common ancestor of urochordates and vertebrates (Wang et al., 2000, 2002; Karabinos et al., 2004) while the type IV proteins are most likely an acquisition of the vertebrate lineages.

Methods

BLAST searches were performed on whole genome data from the NCBI trace archive (http://www.ncbi.nlm.nih.gov) and the draft genome database of the *B. floridae* (http://genome.jgi-psf.org/Braf1). Searches of expressed sequence tags (ESTs) were performed on the NCBI EST database (http://ncbi.nlm.nih.gov/BLAST). Percent identities between different IF sequences were calculated using the program LALIGN at www.ch.embnet.org. The amino acid sequences of the conserved rod domain of various *Branchiostoma* A and B-type IF proteins were aligned using the multiple alignment program ClustalW

and the alignment was submitted to a distance-based unrooted phylogenetic analysis using the unweighted pair-group method with arithmetic mean (UPGMA) or the neighbor-joining method at www.genome.jp. Only 259 amino acid long sequences of the conserved rod domain of various *Branchiostoma*, protostomic and vertebrate IF proteins (Karabinos et al., 2002) were aligned using the multiple alignment program MUSCLE and the alignment was submitted to the unrooted bootstrap (100 replications) maximum likelihood phylogenetic analysis and to the neighbor-joining distance analysis at www.phylogeny.fr. The phylogenies were calculated using the amino acid substitution model Dayhoff. Accession numbers of the new *Branchiostoma* IF sequences, used in the phylogenetic studies, are presented in Table 1; other *Branchiostoma*, protostomic and vertebrate sequences were reported elsewhere (Karabinos et al., 2002).

Acknowledgements

We thank Klaus Weber for discussions. This work was funded by the European Regional Development OPVaV-2009/2.2/05-SORO (ITMS code:26220220143).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejcb.2013.10.004.

References

Erber, A., Riemer, D., Bovenschulte, M., Weber, K., 1998. Molecular phylogeny of metazoan intermediate filament proteins. J. Mol. Evol. 47, 751–762.

Erber, A., Riemer, D., Hofemeister, H., Bovenschulte, M., Stick, R., Panopoulou, G., Lehrach, H., Weber, K., 1999. Characterisation of the Hydra lamin and its gene; a molecular phylogeny of metazoan lamins. J. Mol. Evol. 49, 260–271.

Fuchs, E., Weber, K., 1994. Intermediate filaments: structure, dynamics, function and disease. Annu. Rev. Biochem. 63, 345–382.

Hazes, B., 1996. The $(QxW)_3$ domain: a flexibile lectin scaffold. Prot. Sci. 5, 1490–1501.

Herrmann, H., Hesse, M., Reichenzeller, M., Aebi, U., Magin, T.M., 2003. Functional complexity of intermediate filament cytoskeletons: from structure to assembly to gene ablation. Int. Rev. Cytol. 223, 83–175.

Herrmann, H., Strelkov, S.V., Burkhard, P., Aebi, U., 2009. Intermediate filament: primary determinants of cell architecture and plasticity. J. Clin. Invest. 119, 1772–1783.

Hesse, M., Franz, T., Tamai, Y., Taketo, M.M., Magin, T.M., 2000. Targeted deletion of keratin 18 and 19 leads to trophoblast fragility and early embryonic lethality. EMBO J. 19, 5060–5070.

Hesse, M., Zimek, A., Weber, K., Magin, T.M., 2004. Comprehensive analysis of keratin gene clusters in humans and rodents. Eur. J. Cell Biol. 83, 19–26.

Karabinos, A., Riemer, D., Erber, A., Weber, K., 1998. Homologues of vertebrate type I, II and III intermediate filament (IF) proteins in an invertebrate; the IF multigene family of the cephalochordate *Branchiostoma*. FEBS Lett. 437, 15–18.

Karabinos, A., Riemer, D., Panopoulou, G., Lehrach, H., Weber, K., 2000. Characterisation and tissue-specific expression of the two keratin subfamilies of intermediate filament proteins in the cephalochordate *Branchiostoma*. Eur. J. Cell Biol. 79, 1–10.

Karabinos, A., Wang, J., Wenzel, D., Panopoulou, G., Lehrach, H., Weber, K., 2001a. Developmentally controlled expression pattern of intermediate filament proteins in the cephalochordate *Branchiostoma*. Mech. Dev. 101, 283–288.

Karabinos, A., Schmidt, H., Harborth, J., Schnabel, R., Weber, K., 2001b. An essential role for four intermediate filament proteins in *Caenorhabditis elegans* development. Proc. Natl. Acad. Sci. U.S.A. 98, 7863–7868.

Karabinos, A., Schunemann, J., Parry, D.A.D., Weber, K., 2002. Tissue-specific co-expression and in vitro heteropolymer formation of the two small *Branchiostoma* intermediate filament proteins A3 and B2. J. Mol. Biol. 316, 127–137.

Karabinos, A., Zimek, A., Weber, K., 2004. The genome of the early chordate Ciona intestinalis encodes only five cytoplasmic intermediate filament proteins including a single type I and type II keratin and a unique IF-annexin fusion protein. Gene 326, 123–129.

Karabinos, A., Schunemann, J., Parry, D.A.D., 2012. A rod domain sequence in segment 1B triggers dimerisation of the two small *Branchiostoma* IF proteins B2 and A3. Eur. J. Cell Biol. 91, 800–808.

Li, W., Zhong, J., Wang, Y., 2013. Genetic diversity and population structure of two lancelets along the coast of China. Zool. Sci. 30, 83–91.

Mclean, W.H., Lane, E.B., 1995. Intermediate Filaments in Disease. Curr. Opin. Cell Biol. 7, 118–125.

Parry, D.A.D., Steinert, P.M., 1995. Intermediate Filament Structure. Springer, New York.

Putnam, H.N., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rachevi, M., Shoguchi, E., Terry, A., Yu, J.K., Benito-Gutierrez, E., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J.J., Grigoriev, I.V., Horton, A.C., de Jong, P.J., Jurka, J., Kapitonov, V.V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegaewa, K., Pennacchio, L.A., Salamov, A.A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin, I., Toyoda, T., Bronner-Fraser, A., Fujiyama, M., Holland, A., Holland, L.Z., Satoh, P.W.H., Rokhsar, N.D., 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064–1071.

Riemer, D., Karabinos, A., Weber, K., 1998. Analysis of eight cDNAs and six genes of intermediate filament (IF) proteins in the cephalochordate *Branchiostoma* reveals differences in the IF multigene families of lower chordates and the vertebrates. Gene 211, 361–373.

Riemer, D., Wang, J., Zimek, A., Swalla, J.B., Weber, K., 2000. Tunicates have unusual nuclear lamins with a large deletion in the carboxyterminal tail domain. Gene 255, 317–325.

Vandebergh, W., Bossuyt, F., 2012. Radiation and functional diversification of alpha keratins during early vertebrate evolution. Mol. Biol. Evol. 29, 995–1004.

Vijayaraj, P., Kröger, K., Reuter, U., Windoffer, R., Leube, R.E., Magin, T.M., 2009. Keratins regulate protein biosynthesis through localization of GLUT1 and -3 upstream of AMP kinase and Raptor. J. Cell Biol. 187, 175–184.

Wang, J., Karabinos, A., Schünemann, J., Riemer, D., Weber, K., 2000. The epidermal intermediate filament proteins of tunicates are distant keratins; a polymerisation-competent hetero coiled coil of the *Styela* D protein and Xenopus keratin 8. Eur. J. Cell Biol. 79, 478–487.

Wang, J., Karabinos, A., Zimek, A., Meyer, M., Riemer, D.S., Hudson, C., Lemaire, P., Weber, K., 2002. Cytoplasmic intermediate filament protein expression in tunicate development: a specific marker for the test cells. Eur. J. Cell Biol. 81, 302–311.

Zhang, H., Landmann, F., Zahreddine, H., Rodriguez, D., Koch, M., Labouesse, M., 2011. A tension-induced mechanotransduction pathway promotes epithelial morphogenesis. Nature 471, 99–103.

Zimek, A., Stick, R., Weber, K., 2003. Genes coding for intermediate filament proteins: common features and unexpected differences in the genomes of humans and the teleost fish *Fugu rubripes*. J. Cell Sci. 116, 2295–2302.

Zimek, A., Weber, K., 2005. Terrestrial vertebrates have two keratin gene clusters: striking differences in teleost fish. Eur. J. Cell Biol. 84, 623–635.