Genes coding for intermediate filament proteins: common features and unexpected differences in the genomes of humans and the teleost fish *Fugu rubripes*

Alexander Zimek[1], Reimer Stick[2] and Klaus Weber[1,*]

[1]Max Planck Institute for Biophysical Chemistry, Department of Biochemistry, Am Fassberg 11, D-37077 Goettingen, Germany
[2]Department of Cell Biology, University of Bremen, Leobener Strasse NW II, D-28359 Bremen, Germany
*Author for correspondence (e-mail: office.weber@mpibpc.gwdg.de)

Accepted 24 February 2003
Journal of Cell Science 116, 2295-2302 © 2003 The Company of Biologists Ltd
doi:10.1242/jcs.00444

Summary

We screened the genomic sequences of the teleost fish *Fugu rubripes* for genes that encode cytoplasmic intermediate filament (IF) proteins. Here, we compare the number of genes per subfamily (I to IV) as well as the gene mapping in the human and fish genomes. There are several unexpected differences. *F. rubripes* has a sizeable excess of keratin type I genes over keratin type II genes. Four of the six keratin type II genes map close to four keratin type I genes. Thus, a single keratin II gene cluster (as in mammals) seems excluded. Although a continuous genome sequence is not yet available for *F. rubripes*, it is difficult to see how all 19 keratin type I genes can be collected as in the human genome into a single cluster without the presence of type II genes and various unrelated genes. *F. rubripes* has more type III and type IV genes than humans. Some of the type IV genes acquired additional novel intron positions. One gene even harbors (in addition to the two type IV introns) three novel introns and three introns usually present only in mammalian and *F. rubripes* type I-III genes. This mixture of type IV and type I-III intron positions poses a problem for the traditional view that the first type IV gene arose in evolution by a mRNA-mediated translocation event. In the 42 *F. rubripes* genes analysed here, there are several differences in intron patterns compared with mammalian genes. Most correspond to additional introns in the fish genes. A search for genes encoding nuclear lamins reveals the four established fish lamins (A, B1, B2 and LIII) as well as an unexpected second lamin A.

Key words: *Fugu rubripes*, Gene clusters, Intermediate filaments, Keratin, Lamin, Neurofilament, Vertebrate genomes

Introduction

In humans, the family of genes encoding the structural proteins of the cytoplasmic intermediate filaments (IFs) has more than 60 members and is one of the 100 largest multigene families (Hesse et al., 2001). Sequence identity levels of IF proteins, the organization of the corresponding genes and their expression patterns define several IF subtypes (Fuchs and Weber, 1994; Herrmann and Aebi, 2000; Coulombe et al., 2001). Type I and type II keratins are the largest subfamilies. They give rise to the epithelial keratin filaments that are based on obligate heteromeric double-stranded coiled coils formed by a type I and a type II keratin. Type III includes four proteins that can form homopolymeric IFs. The genes for the seven type IV proteins show an entirely different intron pattern than do type I-III genes. They have only two to three introns related to the central rod domain of the proteins and these introns occur in positions not seen in type I-III genes. The nuclear lamins form the type V, whereas the eye lens proteins filensin and phakinin constitute a separate group (BF, for beaded filaments).

A survey of the draft sequence of the human genome (International Human Genome Sequencing Consortium, 2001) shows that genes coding for non-keratin IF proteins are not clustered (Hesse et al., 2001). By contrast, all type I keratin genes except for K18 form a dense cluster on chromosome 17q21, whereas all type II keratin genes and K18 form a similar cluster on chromosome 12q13 (Waseem et al., 1990).

Point mutations in a still-growing number of IF genes are connected with human diseases. Mutations in at least 14 epidermal keratin genes cause fragility syndromes of the skin (Irvine and McLean, 1999) and similar mutations in the type III desmin gene connect to myopathies of heart and skeletal muscle (Goldfarb et al., 1998), whereas mutations in the GFAP gene are found in Alexander's disease (Brenner et al., 2001; Li et al., 2002). Finally, in *Caenorhabditis elegans*, at least four of the 11 IF genes are essential for nematode development (Karabinos et al., 2001). Type I-III genes are not restricted to vertebrates but have also been documented in the early chordates, which, however, seem to lack type IV genes (reviewed in Karabinos et al., 2002; Wang et al., 2002).

Some genes for type I-IV cytoplasmic IF proteins from fish have previously been documented by cDNA cloning in particular in the goldfish and the rainbow trout (Markl and Schechter, 1998; Schaffeld et al., 2002a,b), and nuclear lamins have been analysed in the goldfish (Yamaguchi et al., 2001)

Table 1. *F. rubripes* IF genes*

Scaffold		Size	Region	Features†	Gap	End‡	IF type§
63	–	268.0	95349-99373		–	–	III: plasticin/peripherin
91	–	234.6	26057-32877		Yes	–	BF: filensin
117	+	211.9	87158-88291		–	–	III: vimentin 2, no introns
120	–	209.3	175776-178130		–	–	IV: X
135	–	203.5	2932-7934	c	–	–	I
137	–	202.8	27765-29473		–	–	IV: NF-L-like 1
138	+	204.8	8429-12857		–	–	III: desmin 1
214	–	173.4	28286-32679	c	–	–	II
214	–	173.4	14539-17560		–	–	II
214	+	173.4	37412-40948		Yes	–	I
285	–	154.8	142763-145343		–	–	II
410	–	133.1	54284-57470		Yes	–	III: GFAP minus-5′
1245	–	70.7	14786-17947		–	–	IV: NF-H-like
1803	–	51.9	50841-51799		–	Yes	I minus-5′
1885	+	51.5	3846-5468		–	–	IV : gefiltin-like 2
1912	+	50.1	38591-40601		–	–	IV : gefiltin/α-internexin
2158	+	44.3	24984-26255		Yes	–	II minus-5′
2158	–	44.3	20141-22916		–	–	I
2208	–	42.7	14518-16881		–	–	IV: gefiltin-like 1
2296	+	42.7	5038-8236		–	–	IV: NF-L-like 2
2330	–	41.1	18142-22707		–	–	III: vimentin 1
2477	+	38.8	22367-29358	c	–	–	IV: Y
2605	–	35.4	20731-24757		Yes	–	I
2605	–	35.4	28812-30882		–	–	I
2605	+	35.4	33974-34970	f,c	Yes	Yes	I minus-3′
2767	+	33.1	9554-12033		–	–	I
2767	+	33.1	1961-4054		–	–	I
2807	+	33.6	23293-25530		–	–	I
2807	–	33.6	14024-16171	c	Yes	–	I
3159	+	30.4	26132-28298		–	–	I
3159	–	30.4	3805-8015		–	–	II: K8
3159	+	30.4	21097-23529	c	Yes	–	I: (K18) minus-3′
3504	+	25.0	24106-24608		–	Yes	III: desmin 2, minus-3′
3830	–	21.3	12182-17002		Yes	–	II minus-3′
4275	–	18.1	14446-17855		–	Yes	III: desmin 2, minus 5′
6593	+	8.9	704-3759	c	Yes	–	IV: NF-M, minus-3′
7203	–	7.6	54-391	f	–	Yes	I minus-3′
7320	–	7.3	1502-3386		–	–	I
7354	–	8.8	895-3688		–	–	I
8680	–	4.9	3781-4544	f,c	Yes	Yes	I minus-5′
8680	+	4.8	1032-2913	f	–	–	I
8762	–	4.8	1221-3118		–	–	I
29247	+	2.4	contig 192-2239		–	Yes	BF: phakinin, minus-5′

*From the database by Aparicio et al. (Aparicio et al. 2002), release 6.1.1.
†Indicates necessary frame shifts (f) and changes from the proposed gene structure (c).
‡Indicates whether or not the gene is at a scaffold end.
§Sequences lacking 5′ or 3′ ends are marked minus-5′ and minus-3′, respectively.

and the zebrafish (Hofemeister et al., 2002). However, only the emerging genome of the teleost fish *Fugu rubripes* (Aparicio et al., 2002) allows a detailed comparison of IF gene organization and complexity in man and a lower vertebrate. Here, we report on some unexpected differences between IF genes in *F. rubripes* and mammals.

Results and Discussion

The *F. rubripes* genome is currently provided by 12,381 non-overlapped scaffolds accounting for a total of 332.5 Mb. Joining of the scaffolds is expected soon (Aparicio et al., 2002). We searched the *F. rubripes* database between October and December 2002. Table 1 summarizes the scaffolds that contain cytoplasmic IF genes. The area number on the scaffold containing the IF gene and its direction are also given. Some genes are incomplete either because of a gap in the sequence or because the gene is located at one end of the scaffold (Table 1). A summary of this search is given in Table 2, which compares the number of *F. rubripes* genes for the different subfamilies with those deduced from a survey made on the draft sequence of the human genome in spring of 2001 (Hesse et al., 2001). The values provided are minimum values and might still increase slightly once the two genomes are completed. In the human genome, there are at least 62 cytoplasmic IF genes. If we subtract the 15 genes encoding hair keratins as a mammalian specialization, this number is reduced to 47. *F. rubripes* has at least 42 genes and thus displays a complexity that is similar to mammals. However, it shows a distinct distribution of the number of genes per subfamilies I-IV (Table 2).

The *F. rubripes* database (Aparicio et al., 2002) is very reliable and well organized. In a few cases we needed to introduce a frame shift to keep the obvious reading frame or to

Table 2. Comparison of IF gene complexity*

Subfamily	*Fugu rubripes*	Human†
Keratin I	19	16+9 h
Keratin II	6	18+6 h
Type III	6	4
Type IV	9	7
BF‡	2	2
Total	42	47+15 h
Lamin V	5	3

*Gene numbers for humans are taken from a previous survey of the human genome draft (Hesse et al., 2001).

†The number of hair keratins is indicated (h); this is a mammalian specialization.

‡Beaded eye lens filament.

explore a major change from the proposed gene structure (Table 1). Two expected difficulties arose: the occasional presence of sequence gaps in some genes situated in the interior of a scaffold and the location of a gene at the end of a scaffold (Table 1). The first problem can be solved directly by PCR amplification bridging the gaps between the known neighboring sequences. The second set of problems, which relates to eight cytoplasmic IF genes, requires overlaps for the more than 12,000 scaffolds, which should be supplied in the future by the *F. rubripes* genome sequencing consortium (Aparicio et al., 2002).

Striking excess of type I over type II keratin genes in *F. rubripes*

Tables 1 and 2 show the presence of 13 complete and six incomplete *F. rubripes* type I genes. The total of 19 genes surpasses the 16 human type I genes (not including the nine type I hair keratin genes thought to be a mammalian specialization). An entirely different situation is given by the type II keratin genes because we located only four complete and two nearly complete genes. Thus, there are about three times as many type I than type II genes in *F. rubripes*, whereas, in humans, the numbers of type I and II genes is similar (Tables 1, 2). The large excess of type I over type II genes could indicate that functional differences between the obligatory heteropolymeric keratin filaments of different cell types depend primarily on the type II genes that are expressed, whereas the type I genes provide additional variability.

Most *F. rubripes* keratin genes show the intron patterns previously described for mammalian type I and type II genes. The two complete keratin I genes on scaffold 2605 have an additional intron between the traditional introns 5 and 6. The keratin II gene on scaffold 3830 has another novel intron position situated between introns 6 and 7. A striking case of an unusual intron pattern is observed in the keratin I gene on scaffold 7354. It has the normal type I intron pattern but, in addition, has an intron that occurs in all mammalian type II and in all *F. rubripes* type II genes (intron 1 of type II genes). The keratin I gene on scaffold 135 shows an unusual doubling of exon 6 (protein sequence identity 95%), which encodes the C-terminal end of the rod domain. Possibly, these exons are alternatives. The type I gene situated at the end of scaffold 8680 (Fig. 1; Table 1) is incomplete and lacks the 5′ end. Interestingly, it is the only *F. rubripes* IF gene that shows several gaps in alignments of the predicted protein sequence and might be a rare pseudogene. Finally the single keratin I genes on the two small scaffolds 7320 and 8762 (Table 1) share 96.8% sequence identity on the nucleotide level including the six introns. This observation is clearly unrelated to the often suggested partial tetraploidity of fish genomes (Aparicio et al., 2002) and instead signals a very recent gene duplication event.

Lack of the keratin II gene cluster in *F. rubripes*

The human genome contains all 25 type I keratin genes except for the keratin 18 gene in a cluster on chromosome 17q21, where they are arranged in the same orientation. Similarly, all 24 type II keratin genes are in a similar cluster on chromosome 12q13, which also harbors the keratin 18 gene next to the keratin 8 gene (Hesse et al., 2001). Although the more than 12,000 scaffolds are not yet arranged as continuous *F. rubripes* genome sequence, the current mapping results (Table 1, Fig. 1) suggest that keratin genes are differently organized in the *F. rubripes* and the mammalian genomes.

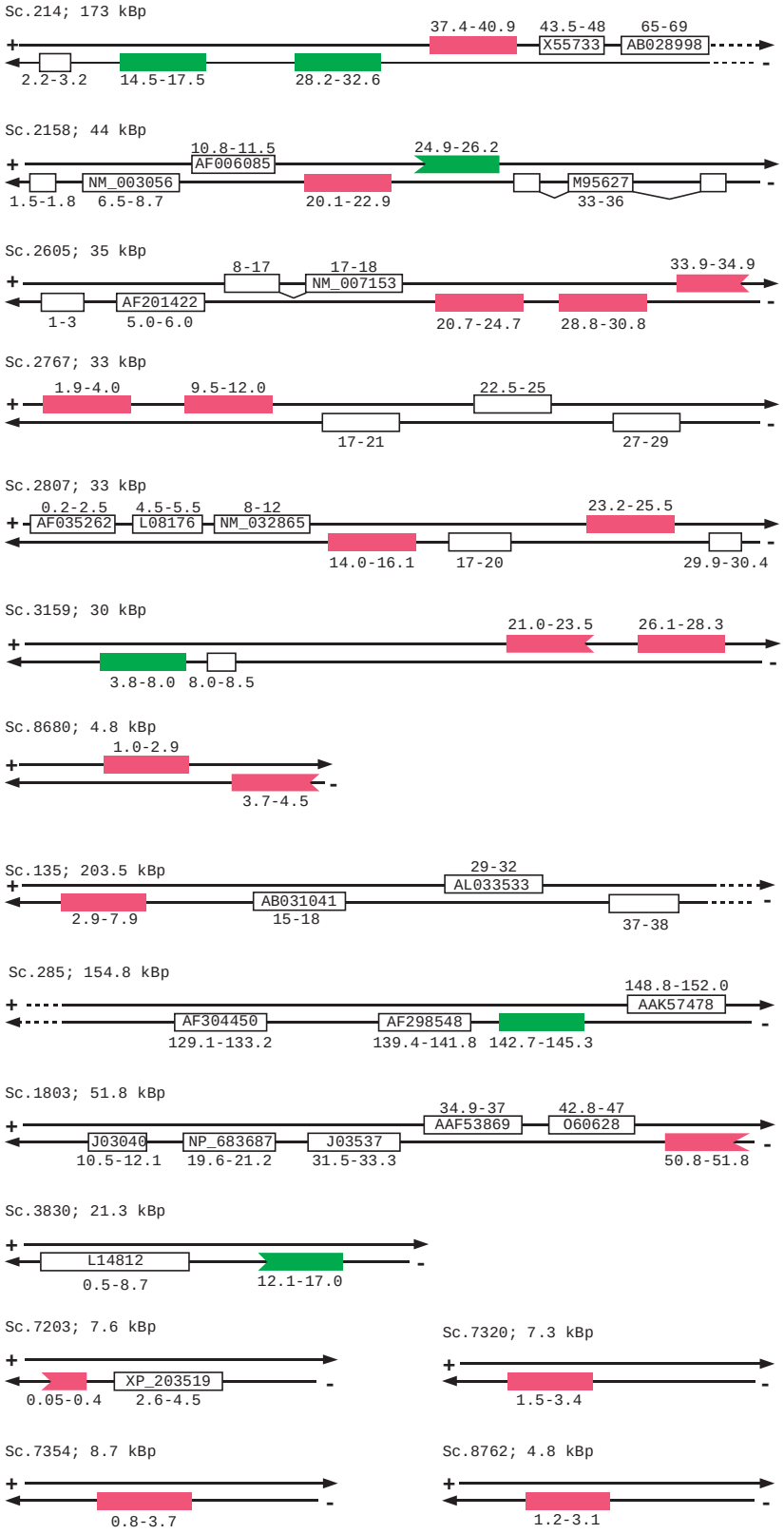
Fig. 1 shows that four of the six type II genes locate as either paired (scaffold 214) or single (scaffolds 2158 and 3159) genes next to either one or two type I genes. The other two type II genes map to two separate scaffolds (285 and 3830). Thus, the presence of a single keratin II gene cluster as in humans is excluded. Interestingly, when type I and II genes map together, they show different orientations.

There is already clear evidence for some clustering of keratin I genes in *F. rubripes*. Fig. 1 shows four groups of two and three directly neighboring keratin I genes, and a pair of keratin I genes separated only by a hypothetical gene. Because another four keratin I genes lie on rather small scaffolds, one could envision a single cluster of many type I genes in *F. rubripes*. However, it seems not to be possible to build a cluster that collects, as in humans, all type I genes without the simultaneous incorporation of four type II genes and various unrelated genes. Neighboring type I genes can have either the same or opposite orientation.

The emerging differences of keratin gene locations in mammalian and fish genomes argue that the keratin gene clustering as documented for mammals was acquired after the fish lineage separated from the lineage leading to higher vertebrates. Forthcoming genomic data on the amphibian *Xenopus tropicalis* and the chicken (as a representative of the birds) will shed light on the question when keratin gene clustering was acquired during vertebrate evolution. We also note that a recent phylogenetic analysis of keratin I and II proteins indicates that fish epidermal keratins diversified independently from the mammalian epidermal keratin radiation, keratins 8 and 18 of interior epithelia are true orthologs in fish and mammals (Schaffeld et al., 2002a).

Mammalian keratins 8 and 18 are typical of internal epithelia and represent the earliest keratin expression pair in embryogenesis. Interestingly the gene for keratin 18, a type I keratin, is adjacent to the keratin 8 gene in the type II gene cluster on human chromosome 12q13 (Waseem et al., 1990; Hesse et al., 2001). The close proximity of keratin 8 and 18 genes also seems to hold for *F. rubripes*. The type II gene on scaffold 3159 codes for a keratin 8 (92% and 78% sequence identity with the rod domain of keratins 8 from rainbow trout

and human, respectively). The keratin 8 gene is separated by a very small hypothetical gene from a type I keratin gene whose sequence is still not complete (Fig. 1). BLAST analysis registers the predicted protein as keratin 18 (81% and 64% identity of the available rod sequence with keratins 18 from rainbow trout and human, respectively). Fig. 2 gives some examples of the sequence similarity of *F. rubripes* and human IF proteins.



Duplication of the type III desmin gene

Previous cDNA cloning studies established in fish the homologs of the four mammalian type III genes encoding vimentin, desmin, GFAP and peripherin [fish peripherin is often referred to as plasticin in the literature] (Markl and Schechter, 1998). The four complete genes are also present in *F. rubripes*, which shows two additional type III genes (Table 1). The fifth type III gene arises by joining scaffolds 3504 and 4275, with the assumption that the scaffold ends protrude into the first intron of the predicted desmin 2 gene. The two gene fragments predict 70-80% sequence identity for desmins 1 and 2. The desmin 2 gene has, in addition to the eight typical type III introns, a further novel intron situated between the traditional introns 5 and 6. Using PCR

Fig. 1. *F. rubripes* scaffolds with two or three keratin genes and the collection of single keratin scaffolds. Scaffolds (Sc), given by number and size, and the areas occupied by the genes are from Table 1. Scaffolds are not drawn to scale. Keratin I and II genes are marked in red and green, respectively. Incomplete genes are marked by an indentation of the colored bars. Notice that four type II genes map next to type I genes, whereas the other two type II genes lie on scaffolds 285 and 3830. Thus, a keratin II gene cluster as in the human genome seems excluded. Also, 13 type I genes occur on seven scaffolds and the other six type I genes lie as single keratin I genes on other scaffolds (Table 1). Although the final view of keratin I gene mapping requires the full *F. rubripes* genome sequence, current results exclude the pure type I cluster present in the human genome. Other genes present on the scaffolds are given as uncolored boxes. Those with a human counterpart carry the corresponding accession number (Aparicio et al., 2002). These read, from top to bottom: X55733, initiation factor 4B; AB028998, tensin 2; NM_003056, solute carrier family (folate transporter), member 1; AF006085, ARP 2/3 complex subunit 2; M95627, angio-associated migratory cell protein AAMP; AF201422, splicing coactivator subunit SRM300; NM_007153, zinc finger protein 208; AF035262, BAF57; L08176, Epstein-Barr-virus-induced G-protein-coupled receptor; NM_032865, tensin-like protein fragment; AL033533, prostaglandin endoperoxide synthase 2; AB031041, LIM-homodomain protein 6.1a; AAK57478, elongation factor 1A binding protein; AF304450, sarcolemma associated protein 1; AF298548, caspase recruitment domain protein 7; AAF53869, *Drosophila* CG13969-PA; 060628, transmembrane 4 superfamily, member 8; J03040, human SPAPC/osteonectin/BM-40; NP_683687, aprataxin, novel nuclear protein; J03537, ribosomal protein S6; L14812, retinoblastoma related protein p107; XP_203519, murine Mycbp-associated protein. Unnamed boxes indicate hypothetical genes.

amplification on *F. rubripes* DNA and sequence analysis, we have verified the proposed arrangement of the two scaffolds covering the desmin 2 gene.

The sixth gene (scaffold 117) has some unexpected features. It is intronless and lies in the large tenth intron (3.2 kb) of a gene encoding the enzyme isoleucine tRNA synthase. The open reading frame predicts a second vimentin that shares ~43% sequence identity with vimentin 1. The canonical sequence motif LNDR in coil 1a is changed to LNAK in vimentin 2. Currently, we do not know whether vimentin 2 is an active gene. The lack of a polyA tract argues against its being a processed pseudogene. The presence of two desmin and vimentin genes in *F. rubripes* and the previous finding of at least two vimentin genes in *Xenopus laevis* (Herrmann et al., 1989), which is a tetraploid species, open the possibility that only higher vertebrates have single vimentin and desmin genes.

Fig. 2. Examples of pairwise sequence comparisons of *F. rubripes* (F) and human (H) type I to IV proteins. (a) A *F. rubripes* keratin I (scaffold 3159) and human keratin 18. (b) A *F. rubripes* keratin II (scaffold 3159) and human keratin 8. (c) *F. rubripes* desmin 1 (scaffold 138) and human desmin (type III). (d) *F. rubripes* and human α-internexin; the *F. rubripes* protein is from scaffold 1912 (Table 1). Identical amino acid residues in each pair are marked by bold print.

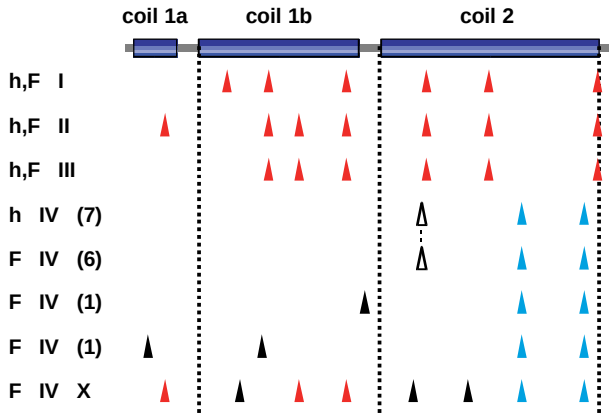


Fig. 3. Schematic diagram of the location of introns in human (h) and *F. rubripes* (F) type I-IV genes. Only introns related to the region encoding the central rod domains (sketch of coils 1a, 1b and 2 at the top) are shown. Notice the related but not identical intron patterns of type I, II and III genes (red triangles), which are shared by human and *F. rubripes*, and the entirely different intron pattern of type IV genes (blue triangles). All seven human type IV genes show the two conserved intron positions, and the NF-H gene has an additional third intron (open triangle) marked by a vertical line. This pattern holds in *F. rubripes* for six genes (gefiltin, gefiltin like-2, NF-L-like 1, NF-M, NF-H-like and Y), with NF-H resembling human NF-H by the additional intron (open triangle marked by a vertical line). The other three *F. rubripes* type IV genes have either one (NF-L-like 2; third line from bottom) or two (gefiltin-like-1; second line from bottom) novel intron positions (black triangles); *F. rubripes* gene X (bottom line) has the two type IV intron positions (blue) and three type I to III intron positions (red) together with three novel intron positions (black).

The beaded filaments of the mammalian and avian eye lens contain the two special cytoplasmic IF proteins, phakinin (CP49) (Hess et al., 1996) and filensin (Masaki and Quinlan, 1997), which together form the BF subfamily. The corresponding genes in *F. rubripes* locate to contig 29247 and scaffold 91, respectively (Table 1). The *F. rubripes* phakinin gene lacks the 5′ end and the consensus sequence at the end of the rod domain of the phakinin protein is, as in other phakinins, changed from YRKLLEGE to YHGILDGE (Sandilands et al., 1998). The *F. rubripes* filensin gene has still a small sequence gap.

Surprisingly many type IV genes

The seven mammalian type IV genes (Lewis and Cowan, 1986) show an entirely different organization than do type I-III genes (Tyner et al., 1985) (Fig. 3). They have only two introns (three for NF-H), which occupy unique positions and occur late in the region encoding the rod domain. To account for this different placement of introns, it was proposed that the first type IV gene arose by an mRNA-mediated transposition event and that subsequent events led to the acquisition of the few new introns (Lewis and Cowan, 1986).

Using the presence of the two intron positions conserved in all mammalian type IV genes, a total of nine type IV genes can be identified in *F. rubripes* (Tables 1, 2). Several of these genes pose problems in annotation compared with the mammalian genes and so some assignments are tentative. The gene on scaffold 1912 predicts a protein with 84% sequence identity to goldfish gefiltin, the fish homolog for mammalian α-internexin (Markl and Schechter, 1998). Indeed, the *F. rubripes* gefiltin-internexin protein shares 60% identity with human internexin (Fig. 2). The gene on scaffold 2208 predicts a gefiltin-like protein (gefiltin-like 1) that shares 74% sequence identity with gefiltin but has a divergent tail domain (identity with human internexin 50%). A further gefiltin like protein, gefiltin-like 2, is coded by the type IV gene on scaffold 1885. Although gefiltin-like 2 shows nearly the same similarity with gefiltin and vimentin over the rod domain, the intron pattern identifies the gene as a type IV gene.

Table 3. *Fugu rubripes* lamin genes

Scaffold		Size	Region	Features*	Gap	End†	Lamin type‡
146	+	198.1	46588-50645		–	–	A
2719	+	36.2	29184-29519		Yes	Yes	A2 minus-3′
4831	–	16.3	4494-12224		–	–	B1
5332	+	13.1	1833-5452	c	–	–	LIII: a and b variants
6482	–	9.4	1-2110		–	Yes	B2 minus-3′
6631	–	9.0	1-7195		Yes	Yes	A2 minus-5′
7678	–	7.5	1-7484	c	Yes	Yes	B2 minus-5′ and minus-3′

*Indicates necessary changes from the proposed gene structure (c).
†Indicates whether or not the gene is at a scaffold end.
‡Sequences lacking 5′ or 3′ ends are marked minus-5′ and minus-3′, respectively.

The two genes present on scaffolds 137 and 2296 predict proteins that are related to gefiltin but have unique tail domains. The second halves of the tail domains are highly acidic owing to the presence of many glutamic acid residues, which often form polyglutamic acid strings. Because this is a distinctive feature of mammalian (Lewis and Cowan, 1986) and *Xenopus* (Charnas et al., 1992) neurofilament triplet NF-L proteins, we tentatively name these *F. rubripes* type IV genes NF-L1 and NF-L2, respectively (Table 1). The *F. rubripes* neurofilament triplet NF-M gene located on scaffold 6593 still has two sequence gaps that obscure the intron pattern. Because of its convincing relation to the corresponding goldfish gene (Glasgow et al., 1994), we used this latter gene in the comparison below. The type IV gene located on scaffold 1245 is tentatively called NF-H because it has the additional intron position of mammalian NF-H genes (Lees et al., 1988) and the predicted protein has a tail domain containing many short degenerate repeats. Depending on the choice between two possible gene structures, there are 19 or 30 degenerate repeats. Whereas the 21 degenerate repeats of mammalian NF-H involve essentially the 14 residue motif KSPEKAKSPVKEEA with two serine phosphorylation sites (Lees et al., 1988), the *F. rubripes* repeat is based on the 10 residue motif ETKPAAKEEP with one threonine phosphorylation site.

Gene Y on scaffold 2477 predicts the only *F. rubripes* type IV protein with a low sequence similarity with gefiltin. The predicted protein has a very small head and a very long tail domain. Although this is a structural feature of mammalian nestin (Lendahl et al., 1990) and synemin (Titeux et al., 2001), no convincing homology was detected. Finally, gene X located on scaffold 120 predicts again a protein of 43% similarity with gefiltin but its astonishing intron pattern (see below) makes an annotation very difficult.

Although the *F. rubripes* collection of type IV genes already includes nine genes, it lacks obvious homologs encoding the large proteins nestin (Lendahl et al., 1990) and synemin (Titeux et al., 2001), the protein syncoilin, which is a constituent IF member of the muscle dystrobrevin complex (Newey et al., 2001). Genes coding for non-keratin IF proteins are not clustered in the human genome (Hesse et al., 2001). Similarly, in *F. rubripes*, there is no scaffold that harbors more than one type III or one type IV gene.

Surprising intron additions and the problem of the origin of type IV genes

Although, in general, fish and mammalian genes have the same intron pattern (Aparicio et al., 2002), some *F. rubripes* type IV genes do not (Fig. 3). The genes for gefiltin, gefiltin-like 2, NF-L1, NF-M and protein Y have only the conserved two intron positions of mammalian type IV proteins. An additional intron is found in the same position in mammalian and *F. rubripes* NF-H genes. However, the genes encoding NF-L2 and gefiltin-like 1 have one or two additional introns situated at novel positions. Even more complex is the situation in gene X, which has eight introns: the two conserved intron positions of type IV genes, three novel intron positions and a further three positions that are characteristically found only in mammalian and *F. rubripes* type I-III genes. These include the first intron position of type II genes, the third intron position of type II genes (which is also present in type III genes) and the intron position corresponding to the end of the coil 1b domain, which is found in all type I-III genes. The documentation of a fish IF gene that combines type I-III intron positions with type IV intron positions (Fig. 3) is at first difficult to accommodate in a model assuming that the first type IV gene arose by translocation of an intronless mRNA into the genome (Lewis and Cowan, 1986). One possibility for the origin of this gene X that stays within this model is the speculation that it arose as a chimera of a keratin II and a type IV gene (Fig. 3).

Genes NF-L2, gefiltin-like 1 and X together provide a total of six new intron positions of IF genes that have no counterpart in human IF genes. The number of novel fish IF intron positions is increased to ten by the novel intron positions in two type I keratin genes, one type II gene and the desmin 2 gene (see above). If we consider vimentin 2 as a special case, there are ten intron gains in the *F. rubripes* IF genes analysed, but no unusual intron loss (except for vimentin 2).

F. rubripes has two A lamins

Previous studies have shown that fish have four nuclear lamins. Whereas lamins A, B1 and B2 are found in all classes of vertebrates, the additional lamin LIII is only detected in amphibia and fish (Döring and Stick, 1990; Yamaguchi et al., 2001; Hofemeister et al., 2002). Table 3 shows that the genomic *F. rubripes* sequences cover the complete genes for lamins A, B1 and LIII (with its two alternative last exons, which produce the isoforms LIIIa and LIIIb). The intron pattern of these three genes is perfectly conserved between fish and human. The lamin B2 gene bridges scaffolds 6482 and 7682. Exon 1 is located to scaffold 6482, where it is followed by a long intron sequence that overlaps extensively with the end of scaffold 7678. This scaffold carries also the middle part of the gene, but the 3′ end is probably obscured by the following large sequence gap. Interestingly, the *F. rubripes* B2 lamin gene has an additional intron inserted in the region encoding the coil 2a domain. Unexpectedly, a second lamin A is also indicated (Table 3). Lamin A2 starts in scaffold 2719 (exon 1 plus intron) and continues with scaffold 6631 (exon 2 till end). Using the zebrafish lamin A as reference (Hofemeister et al., 2002), we find sequence identity of 70% for both *F. rubripes* A lamins, which share only 63% identity. This is the first report of the presence of two lamin A genes in a vertebrate genome. It raises the question of whether one of them belongs to those genes that contribute to the partial tetraploidy of *F. rubripes* (Aparicio et al., 2002). The comparatively low degree of sequence similarity would indicate an ancient duplication event.

We thank M. Osborn (Goettingen) and M. Hesse (Bonn) for helpful discussions.

References

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A. et al. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301-1310.

Brenner, M., Johnson, A. B., Boespflug-Tanguy, O., Rodriguez, D., Goldman, J. E. and Messing, A. (2001). Mutations in GFAP, encoding glial fibrillary acidic protein, are associated with Alexander disease. *Nat. Genet.* **27**, 117-120.

Charnas, L. R., Szaro, B. G. and Gainer, H. (1992). Identification and developmental expression of a novel low molecular weight neuronal intermediate filament protein expressed in *Xenopus laevis*. *J. Neurosci.* **12**, 3010-3024.

Coulombe, P. A., Ma, L. L., Yamada, S. and Wawersik, M. (2001). Intermediate filaments at a glance. *J. Cell Sci.* **114**, 4345-4347.

Döring, V. and Stick, R. (1990). Gene structure of nuclear lamin LIII of *Xenopus laevis*: a model for the evolution of IF proteins from a lamin-like ancestor. *EMBO J.* **9**, 4073-4081.

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

Glasgow, E., Hall, C. M. and Schechter, N. (1994). Organization, sequence, and expression of a gene encoding goldfish neurofilament medium protein. *J. Neurochem.* **63**, 52-61.

Goldfarb, L. G., Park, K. Y., Cervenakova, L., Gorokhova, S., Lee, H. S., Vasconcelos, O., Nagle, J. W., Semino-Mora, C., Sivakumar, K. and Dalakas, M. C. (1998). Missense mutations in desmin associated with familial cardiac and skeletal myopathy. *Nat. Genet.* **19**, 402-403.

Herrmann, H. and Aebi, U. (2000). Intermediate filaments and their associates: multi-talented structural elements specifying cytoarchitecture and cytodynamics. *Curr. Opin. Cell Biol.* **12**, 79-90.

Herrmann, H., Fouquet, B. and Franke, W. W. (1989). Expression of intermediate filament proteins during development of *Xenopus laevis*. I. cDNA clones encoding different forms of vimentin. *Development* **105**, 279-298.

Hess, J. F., Casselman, J. T. and FitzGerald, P. G. (1996). Gene structure and cDNA sequence identify the beaded filament protein CP49 as a highly divergent type I intermediate filament protein. *J. Biol. Chem.* **271**, 6729-6735.

Hesse, M., Magin, T. M. and Weber, K. (2001). Genes for intermediate filament proteins and the draft sequence of the human genome: novel keratin genes and a surprisingly high number of pseudogenes related to keratin genes 8 and 18. *J. Cell Sci.* **114**, 2569-2575.

Hofemeister, H., Kuhn, C., Franke, W. W., Weber, K. and Stick, R. (2002). Conservation of the gene structure and membrane targeting signals of germ cell specific lamin LIII in amphibians and fish. *Eur. J. Cell Biol.* **81**, 51-60.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Irvine, A. D. and McLean, W. H. (1999). Human keratin diseases: the increasing spectrum of disease and subtlety of the phenotype-genotype correlation. *Br. J. Dermatol.* **140**, 815-828.

Karabinos, A., Schmidt, H., Harborth, J., Schnabel, R. and Weber, K.

(2001). Essential roles for four cytoplasmic intermediate filament proteins in *Caenorhabditis elegans* development. *Proc. Natl. Acad. Sci. USA* **98**, 7863-7868.

Karabinos, A., Schunemann, J., Parry, D. A. and 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.

Lees, J. F., Shneidman, P. S., Skuntz, S. F., Carden, M. J. and Lazzarini, R. A. (1988). The structure and organization of the human heavy neurofilament subunit (NF-H) and the gene encoding it. *EMBO J.* **7**, 1947-1955.

Lendahl, U., Zimmerman, L. B. and McKay, R. D. (1990). CNS stem cells express a new class of intermediate filament protein. *Cell* **60**, 585-595.

Lewis, S. A. and Cowan, N. J. (1986). Anomalous placement of introns in a member of the intermediate filament multigene family: an evolutionary conundrum. *Mol. Cell. Biol.* **6**, 1529-1534.

Li, R., Messing, A., Goldman, J. E. and Brenner, M. (2002). GFAP mutations in Alexander disease. *Int. J. Dev. Neurosci.* **20**, 259-268.

Markl, J. and Schechter, N. (1998). Fish intermediate filament proteins in structure, function and evolution. In *Intermediate Filaments* (ed. H. Hermann and J. R. Harris), pp. 1-33. New York: Plenum Press.

Masaki, S. and Quinlan, R. A. (1997). Gene structure and sequence comparisons of the eye lens specific protein, filensin, from rat and mouse: implications for protein classification and assembly. *Gene* **201**, 11-20.

Newey, S. E., Howman, E. V., Ponting, C. P., Benson, M. A., Nawrotzki, R., Loh, N. Y., Davies, K. E. and Blake, D. J. (2001). Syncoilin, a novel member of the intermediate filament superfamily that interacts with alpha-dystrobrevin in skeletal muscle. *J. Biol. Chem.* **276**, 6645-6655.

Sandilands, A., Masaki, S. and Quinlan, R. A. (1998). Lens intermediate filament proteins. In *Intermediate Filaments* (ed. H. Hermann and J. R. Harris), pp. 291-318. New York: Plenum Press.

Schaffeld, M., Höffling, S., Haberkamp, Conrad, M. and Markl, J. (2002a). Type I keratin cDNAs from the rainbow trout: independent radiation of keratins in fish. *Differentiation* **70**, 282-291.

Schaffeld, M., Haberkamp, M., Braziulis, E., Lieb, B. and Markl, J. (2002b). Type II keratin cDNAs from the rainbow trout: implications for keratin evolution. *Differentiation* **70**, 292-299.

Titeux, M., Brocheriou, V., Xue, Z., Gao, J., Pellissier, J. F., Guicheney, P., Paulin, D. and Li, Z. (2001). Human synemin gene generates splice variants encoding two distinct intermediate filament proteins. *Eur. J. Biochem.* **268**, 6435-6449.

Tyner, A. L., Eichman, M. J. and Fuchs, E. (1985). The sequence of a type II keratin gene expressed in human skin: conservation of structure among all intermediate filament genes. *Proc. Natl. Acad. Sci. USA* **82**, 4683-4687.

Wang, J., Karabinos, A., Zimek, A., Mayer, M., Riemer, D., Hudson, C., Lemaire, P. and 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.

Waseem, A., Gough, A. C., Spurr, N. K. and Lane, E. B. (1990). Localization of the gene for human simple epithelial keratin 18 to chromosome 12 using polymerase chain reaction. *Genomics* **7**, 188-194.

Yamaguchi, A., Yamashita, M., Yoshikuni, M. and Hagahama, Y. (2001). Identification and molecular cloning of germinal vesicle lamin B3 in goldfish (*Crassius auratus*) oocytes. *Eur. J. Biochem.* **268**, 932-939.