Available online at www.sciencedirect.com
SCIENCE DIRECT
Genomics 87 (2006) 265 – 274
GENOMICS
www.elsevier.com/locate/ygeno

Phylogenetic analysis of three complete gap junction gene families reveals lineage-specific duplications and highly supported gene classes

Stephen D. Eastman [a], Tim H.-P. Chen [b], Matthias M. Falk [a], Tamra C. Mendelson [a], M. Kathryn Iovine [a,*]

[a] *Department of Biological Sciences, Lehigh University, 111 Research Drive, Iacocca B-217, Bethlehem, PA 18015, USA*
[b] *Genetics Department, School of Medicine, Washington University, St. Louis, MO 63110, USA*

Received 29 July 2005; accepted 17 October 2005
Available online 7 December 2005

Abstract

Gap junctions, composed of connexin proteins in chordates, are the most ubiquitous form of intercellular communication. Complete connexin gene families have been identified from human (20) and mouse (19), revealing significant diversity in gap junction channels. We searched current databases and identified 37 putative zebrafish connexin genes, almost twice the number found in mammals. Phylogenetic comparison of entire connexin gene families from human, mouse, and zebrafish revealed 23 zebrafish relatives of 16 mammalian connexins, and 14 connexins apparently unique to zebrafish. We found evidence for duplication events in all genomes, as well as evidence for recent tandem duplication events in the zebrafish, indicating that the complexity of the connexin family is growing. The identification of a third complete connexin gene family provides novel insight into the evolution of connexins, and sheds light into the phenotypic evolution of intercellular communication via gap junctions.

Keywords: Zebrafish; Connexin; Gap junction; Phylogeny; Gene family

Introduction

Connexins are integral membrane proteins that oligomerize to form gap junctions, proteinaceous channels that permit the transfer of small molecules (<1 kDa) between neighboring cells [1]. Cell–cell communication via gap junctions is critical for normal cellular function and homeostasis as evidenced by the wide variety of connexin mutations that lead to human disease [2]. A single connexin protein folds into four conserved transmembrane domains, one cytoplasmic loop, two extracellular loops, and cytoplasmic amino- and carboxy-termini [3]. Six connexins form a connexon (or hemichannel), and a single gap junction channel forms when two connexons from adjacent cells dock at the plasma membrane. All connexin genes have likely been identified from the human (n = 20) and mouse (n = 19) genomes, revealing large gene families [4]. Why such a diversity of connexin proteins would be required for this seemingly simple function is not yet clear, but one favorable hypothesis is that gap junctional communication is influenced by the composition of channels [2]. Gap junctions may be composed of multiple combinations of connexin isoforms leading to differences in pore size, charge specificities, and gating properties [5–7]. The tissue-specific complement of connexin genes is likely responsible for the precise regulation of gap junctional communication, suggesting that the complexity of the connexin family contributes to the specialization of intercellular communication among different tissues.

Here we describe what appears to be the entire connexin gene family for the zebrafish, *Danio rerio*. A systematic search of genomic databases revealed a remarkable 37 zebrafish connexin genes, the largest connexin gene family yet described and nearly twice the number found in human and mouse. A phylogenetic analysis of the three complete gene families (i.e., human, mouse, and zebrafish) indicates that the large number of zebrafish connexins is not strictly due to a whole-genome duplication event hypothesized to have occurred in the teleost lineage [8]. Rather, some mammalian connexins are absent

* Corresponding author. Fax: +1 610 758 4004.
E-mail address: mki3@lehigh.edu (M.K. Iovine).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.10.005

Table 1
Zebrafish connexin genes

Zebrafish connexin	Previous name	Predicted MW	Human ortholog	Closest human connexin relative	BAC	Position on BAC	Plus/Minus strand	Exons in coding region	Length	Accession #	Reference
cx43.4	n.a.	43.52	n.a.	novel	zK79P17	51,205–52,347	Minus	2	380 aa	NM_131069, X96712.1 ENSDARG00000007099 L46801.1, BC044357.1 NM_131069.2, *BQ169321.1* *CB357124.1*	Essner et al., 1996 Desplantez et al., 2003
cx44.2	n.a.	44.14	n.a.	novel	zK235F19	61,507–62,668	Minus	2	391 aa	NM_131810.2, AF072451.1 BC045279 *ENSDARG00000034207*	n.a.
cx44.6	n.a.	44.56	n.a.	novel	zC203N19.00686 zK235F19 zK91F15	11,976–13,142 10,251–11,412 9041–10,207	Plus Minus Plus	2 2 2	394 aa 394 aa 394 aa	*ENSDARG00000034129* *ENSDARG00000034867* *ENSDARG00000030223*	n.a.
cx45.1	n.a.	45.08	n.a.	novel	zK235F19	45,413–46,574	Minus	2	399 aa	*ENSDARG00000029603*	n.a.
cx47.1	n.a.	47.13	n.a.	novel	n.a.	n.a.	n.a.	n.a.	409 aa	NM_001004574.1 BC081653.1 *ENSDARG00000006961*	n.a.
cx52.8	n.a.	52.84	*CX45/GJA7*	*CX45/GJA7*	zK22A20 zC10J9	59,675–61,108 141,383–142,816	Minus Minus	1 1	477 aa 477 aa	*ENSDARG00000002681*	n.a.
cx27.5	n.a.	27.52	*CX32/GJB1*	*CX32/GJB1*	zK110K5	128,078–131,250	Minus	1	240 aa	ENSDARG00000035553 AF304049.1, BC056546.1 NM_131811.2, *AF028723.1*	Dermietzel et al., 2000.
cx28.6	n.a.	28.58	n.a.	novel	n.a.	n.a.	n.a.	n.a.	251 aa	ENSDARG00000003925 NM_001007212.1, AY135443.1	n.a.
cx28.8	n.a.	28.81	n.a.	*CX25/GJB7*	dZ189A20	26,422–27,186	Plus	1	254 aa	BX511077, AL954305	n.a.
cx30.9	n.a.	31.01	n.a.	novel	zC51F17	107,945–108,748	Plus	1	267 aa	NM_001007288.1 BC077156.1, *AF028724.1* *ENSDARG00000028833*	n.a.
cx31.7	n.a.	31.74	*CX32/GJB1*	*CX32/GJB1*	zC125C23	63,393–64,220	Minus	1	275 aa	n.a.	n.a.
cx33.8	n.a.	30.33	n.a.	*CX26/GJB2* *CX30/GJB6*	zK51D17	88,794–89,597	Plus	1	267 aa	BC095350, NM_212825	n.a.
cx34.4	n.a.	34.41	n.a.	*CX30.3/GJB4* *CX31.1/GJB5*	zC195O20 zK74H17	100,131–101,030 191–1090	Minus Minus	1 1	299 aa 299 aa	AY135444.1, NM_207169 *ENSDARG00000009863*	n.a.
cx35.4	n.a.	35.41	*CX31/GJB3*	*CX31/GJB3*	zC195O20	91,622–92,536	Minus	1	304 aa	BC093154.1 *ENSDARG00000002087*	n.a.
cx52.6	n.a.	52.62	*CX62*	*CX62*	zKp117E10	58,685–60,085	Plus	1	466 aa	NM_212819.1, BX510339 AF465750.1 ENSDARG00000034930	Hussain et al., 2003 Zoidl et al., 2004
cx52.7	n.a.	52.67	*CX62*	*CX62*	n.a.	n.a.	n.a.	n.a.	464 aa	CAAK01006521	n.a.
cx52.9	n.a.	52.89	*CX59/GJA10*	*CX59/GJA10*	n.a.	n.a.	n.a.	n.a.	473 aa	NM_207093, BC066457 *ENSDARG00000002232*	n.a.
cx55.5	n.a.	55.48	*CX59/GJA10*	*CX59/GJA10*	zK66C10 zC9M16.00502 zC177P5	67,015–65,519 55,172–56,668 88,474–87,037	Minus Minus Minus	1 1 1	498 aa 498 aa 498 aa	NM_131812.1 AF304048.1 ENSDARG00000031820 ENSDARG00000004403	Dermietzel et al., 2000 Hussain et al., 2003

cx28.1	cx20d	28.15	n.a.	novel	zK261A18	18,239–18,982	Minus	1	247 aa	CR352322.5	Iovine et al., 2005
cx28.9	cx20c	28.9	n.a.	novel	zK261A18	11,866–12,627	Plus	1	253 aa	NM_001007324.1 BC085626.1, CR352322.5	Iovine et al., 2005
cx32.2	cx20e	31.89	n.a.	novel	zK261A18	33,831–34,664	Plus	1	277 aa	CR352322.5 ENSDARG00000021696	Iovine et al., 2005 Chatterjee et al., 2005
cx32.3	cx20b	32.34	n.a.	novel	zK261A18	4182–5039	Plus	1	285 aa	NM_199612.1, BC054589.1 CR352322.5 ENSDARG00000024664	Iovine et al., 2005
cx34.5	cx20a	34.54	n.a.	novel	zK261A18	556–1458	Plus	1	298 aa	CR352322.5	Iovine et al., 2005
cx39.4	n.a.	39.41	n.a.	novel	zC261O1	14,805–15,830	Minus	1	341 aa	*ENSDARG00000009593*	n.a.
cx39.9	cx46l	39.91	n.a.	*CX46/GJA3*	zK110K5	119,965–121,026	Minus	1	353 aa	NM_212826, AY135445.1 ENSDARG00000004082 *CB363073, CB363241*	Iovine et al., 2005
cx40.8	cx43l	40.79	n.a.	*CX43/GJA1*	n.a.	n.a.	n.a.	n.a.	354 aa	ENSDARG00000007288	Iovine et al., 2005
cx41.8	n.a.	41.81	*CX40/GJA5*	*CX40/GJA5*	dZ225A24	66,320–67,432	Plus	1	370 aa	ENSDARG00000020587	n.a.
cx43	cx43.3	43.46	*CX43/GJA1*	*CX43/GJA1*	zK261A18	41,046–42,191	Plus	1	381 aa	CR352322.5, AY340236.1 ENSDARG00000012589 AY313942.1, NM_131038.1 AF035481.1, *AF067407.1* *BC049297.1*	Cheng et al., 2003 Iovine et al., 2005 Chatterjee et al., 2005
cx44.1	cx44.2	44.1	*CX50/GJA8*	*CX50/GJA8*	dZ225A24	34,670–35,845	Minus	1	391 aa	NM_131809.2, AF288817.1 AF304050.1 *ENSDARG00000015076*	Dermietzel et al., 2000 Cason et al., 2001 Cheng et al., 2003
cx45.6	n.a.	45.6	*CX40/GJA5*	*CX40/GJA5*	zK19H21.02753 zK19H21.04392	185,658–186,860 185,713–186,915	Minus Plus	1 1	400 aa 400 aa	AF531762, NM_001007213 ENSDARG00000001989	Christie et al., 2004
cx48.5	n.a.	48.5	*CX46/GJA3*	*CX46/GJA3*	zK51D17	93,828–95,132	Plus	1	434 aa	ENSDARG00000021889 NM_207642.1, AF520815.1 AF465751.1	Cheng et al., 2003 Cheng et al., 2004
cx50.5	n.a.	50.52	*CX50/GJA8*	*CX50/GJA8*	zK18P12	59,360–61,489	Plus	1	444 aa	*ENSDARG00000007859*	n.a.
cx34.1	n.a.	34.15	n.a.	*CX36/GJA9*	zC46E8 zC87N9	57,199–79,438 82,265–105,107	Plus Plus	2 2	299 aa 299 aa	*ENSDARG00000032464* ENSDARG00000031758 *ENSDARG00000014503*	n.a.
cx35	n.a.	35.07	*CX36/GJA9*	*CX36/GJA9*	zC120J23.01395 bZ1C22 zC260D9	45,973–48,836 93,015–95,878 91,729–94,622	Minus Minus Minus	2 2 2	304 aa 304 aa 304 aa	AF462040.1, AF512548.1 NM_194420.1 ENSDARG00000031311 ENSDARG00000034516 ENSDARG00000034334	McLachlan et al., 2003 Valiunas et al., 2004
cx35.8	n.a.	>35.85	n.a.	*CX36/GJA9*	n.a.	n.a.	n.a.	n.a.	316+ aa	*ENSDARG00000017927*	n.a.
cx40.5	n.a.	40.48	n.a.	novel	zC244E12	51,853–52,917	Plus	1	354 aa	*ENSDARG00000009334*	n.a.
cx46.8	n.a.	46.78	*CX40.1*	*CX40.1*	zK18N12 zK261I18.00385	23,460–26,681 85,822–87,028	Minus Plus	2 2	422 aa 422 aa	n.a.	n.a.

Connexin genes are arranged in clades identifiable by similar colors used on phylogram tree (Fig. 1). Zebrafish connexin genes are subsequently organized by ascending molecular weight within a clade. “Previous name” indicates previous publication name or annotated name. The zebrafish connexin with closest human orthology is indicated under “Human ortholog”. Zebrafish connexins that are closely related to a human connexin are indicated under “Closest human connexin relative”. “Position on BAC”, “Plus/Minus strand”, “Exons in coding region” refers to the BAC located on the same row under “BAC”. “Accession #” lists all gene accession numbers for each connexin. Italicized accession numbers indicate partial predicted transcripts. Boldfaced entries are the gene names for the zebrafish connexins.

from the zebrafish genome, some are found as single relatives, and others are found in multiple copies. In addition, the zebrafish has 14 apparently novel connexins, several of which arose by recent tandem duplication events. This analysis provides evidence that the connexin gene family is increasing in complexity within independently evolving lineages, potentially leading to lineage-specific specialization of gap junctional communication. The evolution of this large gene family may therefore contribute to the development of increasingly complex and diverse cellular functions.

Mammalian connexin genes are named based on their separation into classes (α, β, γ) using the prefix "GJ" for "gap junction" (i.e., *GJA1* for the first member of the α class [9,10]), whereas proteins are named for their differences in size using the prefix "Cx" followed by the predicted molecular weight (i.e., Cx43 [11]). However, classification of some mammalian connexins has been ambiguous due to lack of a single criterion for this purpose. Our phylogenetic analysis, which includes 76 connexins from human, mouse, and zebrafish, identifies the α, β, and γ classes (as well as a potential fourth class) as largely monophyletic, highly supported clades on a phylogenetic tree. Indeed, the inclusion of a distantly related connexin gene family with the mammalian connexins validates the use of clades to define the connexin classes and provides strong evidence that connexin classes are common to all vertebrates. This analysis facilitates further investigation of zebrafish and mammalian connexins by providing a broad, comparative perspective for examining the evolutionary history of the connexin gene family.

Results and discussion

Identification of 37 putative connexin genes in the zebrafish genome

Sixteen zebrafish connexins have been reported in the literature (Table 1) [14–19,22,34–39]. To discover additional zebrafish connexins, we first compared the nucleotide sequence of the 16 reported zebrafish connexins with the whole-genome shotgun (WGS) assembly sequence, version 5, via Ensembl (http://www.ensembl.org/Danio_rerio/). A search of the zebrafish WGS assembly (~96% complete) revealed an additional 18 putative connexins (Table 1). Next, we compared these 34 zebrafish connexin sequences to the finished and unfinished genomic BAC sequence database (http://www.sanger.ac.uk/cgi-bin/blast/submitblast/d_rerio), where approximately 50% of the genome is available as highly reliable contiguous sequence. This revealed the BAC location for 15 of the 16 previously reported connexins (i.e., >95% identity to query sequences), 12 of the 18 connexins identified from the WGS search (i.e., >95% identity to query sequences), as well as 5 additional sequences (i.e., 60–93% identity to query sequences), bringing the total number of putative zebrafish connexins up to 39. Next, we completed similar searches of the trace file database (associated with the WGS project, http://www.ncbi.nlm.nih.gov/genome/seq/DrBlast.html) and identified two additional sequences representing partial connexins. Examination of the electropherograms representing each trace file (http://trace.ensembl.org/perl/traceview) revealed that one of these sequences is reliable (i.e., high-quality sequence data, zDH52-90d08.p1k) and one is unreliable (i.e., low-quality sequence data, zfish44908-752d05.p1k). Future assemblies should identify sequences that overlap with the former trace file and determine whether this represents an additional connexin. In contrast, it is likely that the poor quality of the latter trace file did not permit its assembly with other high-quality sequences and therefore may not represent a new connexin. Since the status of both of these sequences is questionable, we do not include the trace file data in further analyses or in our total count of zebrafish connexins.

Thirty-seven of the 39 connexins exhibit all of the criteria for connexin proteins (described in [3]) including significant sequence homology in each of the four transmembrane domains, an amphipathic motif in the third transmembrane domain, and the three conserved cysteine residues in each extracellular loop. The remaining 2 of the 39 identified genes may not represent connexins since the predicted polypeptides contain only 2 conserved cysteine residues per extracellular loop (i.e., NCBI gene NM_001013546 or BC091468 identified from BAC zK283F18, position 34438–36613; zC159A3.00872, position 39173–43517). Because we were unable to classify these genes as connexins, their sequences were excluded from further analyses.

Finally, we compared the 37 putative connexin sequences with the zebrafish EST database (via NCBI blastn, http://www.ncbi.nlm.nih.gov/) and with the NCBI gene database. Neither search yielded additional connexin sequences. However, the EST database contains one or more EST sequences for 21 of the 37 zebrafish connexins (see Supplementary Material) and the NCBI gene database contains complete mRNA sequences for 20 of the 37 connexins (3 in addition to the EST database; Table 1), providing evidence that at least 24 zebrafish connexins are expressed. The remaining 13 may be expressed at lower levels, at specific stages of development, or in tissues not represented in the current cDNA libraries.

Zebrafish connexins were previously named using the prefix "Cx" followed by the predicted molecular weight of the protein (Table 1). In accordance with this precedent and with the Zebrafish Nomenclature Committee (http://www.grs.nig.ac.jp:6070/), we named the remaining zebrafish connexins similarly using "Cx" when designating proteins and "*cx*" when designating genes. To avoid confusion, non-orthologous zebrafish connexin genes were not given a name already assigned to a human or mouse connexin.

Orthologous and novel zebrafish connexins

We based orthology assignments on identified relationships from the phylogenetic tree (Fig. 1). To be considered orthologous, the zebrafish orthologue must be the closest relative to the mammalian connexin, and the orthologous relationship must be strongly supported (bootstrap >95%). Based on these criteria, we identified 16 zebrafish connexins as having human orthologues. Phylogenetic analysis revealed 11

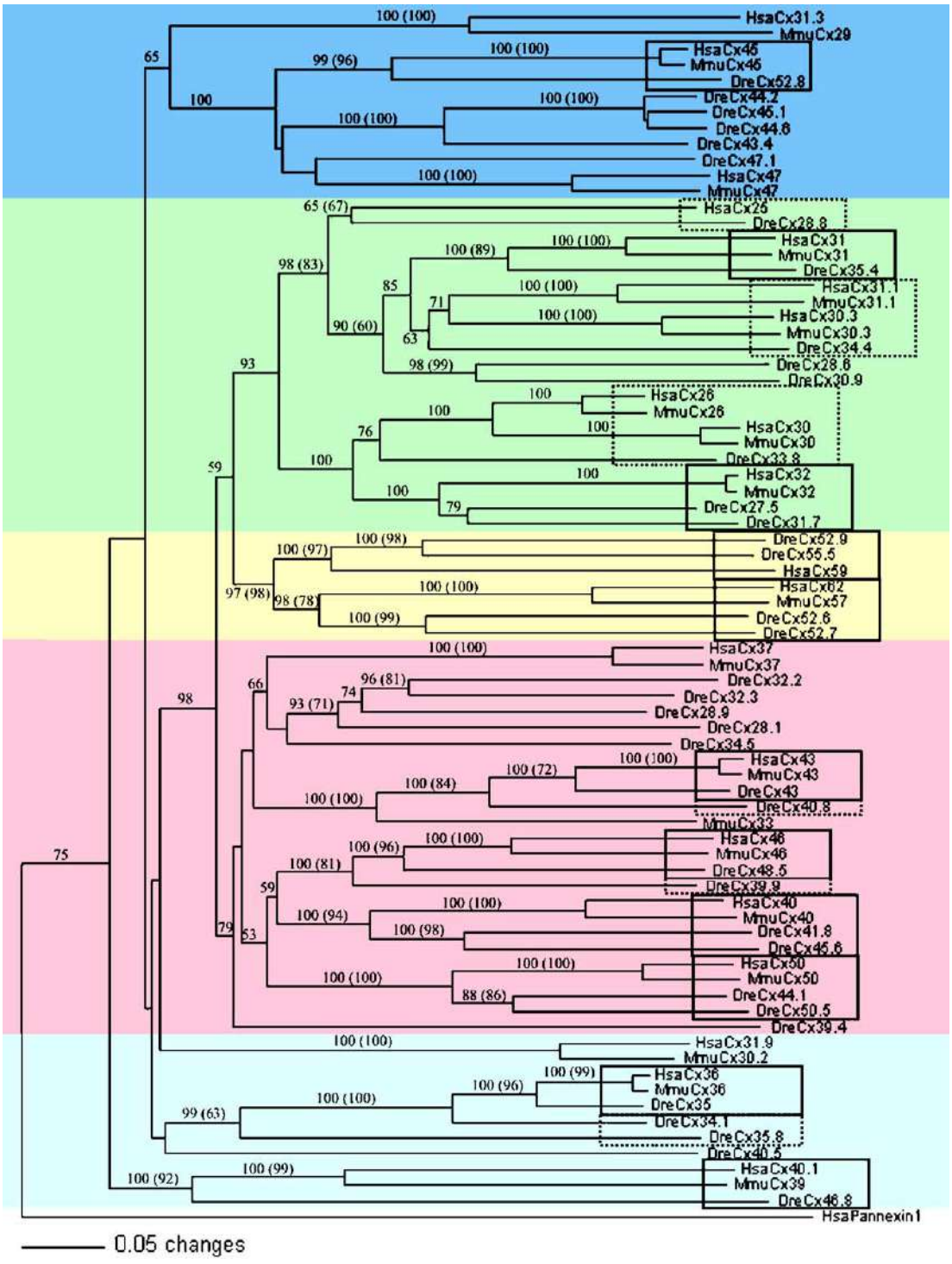

Fig. 1. Phylogeny comparing human, mouse, and zebrafish connexins. A neighbor-joining distance tree comparing full-length amino acid sequences is shown. Bootstrap values >50% are shown for the neighbor-joining analysis and the parsimony analysis (in parentheses). Orthologous relationships are indicated by solid boxes; closely related (and non-orthologous) relationships are indicated by dashed boxes. Major clades are distinguished by colors: dark blue (containing γ-type connexins), green (containing β-type connexins), yellow (containing members of a potential new class), pink (containing α-type connexins), and light blue (singleton connexins). Hsa, *Homo sapiens*; Mmu, *Mus musculus*; Dre, *Danio rerio*.

strongly supported clades (bootstrap values >95%) that each contained 1 human connexin plus 1 (n = 6) or 2 (n = 5) zebrafish connexins (indicated by solid boxes in Fig. 1). Human connexins in the clades containing 2 zebrafish connexins were identified as orthologues only if the 2 zebrafish connexins were each other's closest relatives, suggesting a recent duplication event in the zebrafish lineage. Ten of these 11 groups contained a single mouse connexin as orthologue to the human connexin. Indeed, our results were consistent with all previously identified human and mouse orthologous pairs. We identified 7 additional zebrafish connexins as being closely related to human connexins (indicated by dashed boxes in Fig.

1), though not strictly orthologous as they did not meet the criteria described above. Although we could identify a closest human relative to these connexins, 4 of these were less closely related to their human counterpart than was 1 of the aforementioned 16 (i.e., Cx40.8, Cx39.9, Cx34.1, and Cx35.8), 2 were equally related to 2 human connexins, and in one case (human Cx25, zebrafish Cx28.8), the relationship was not strongly supported (bootstrap <75%). Fourteen zebrafish connexin sequences had no clear human orthologues. However, others have examined conservation of synteny to facilitate the identification of orthologous relationships between zebrafish and human genes [12]. To determine whether any of the 14 novel connexins might represent orthologues based on their syntenic relationships to human connexins, we searched for conservation of synteny. No evidence for synteny of the 14 novel zebrafish connexins was found (data not shown).

Comparing the topologies of several phylogenetic reconstructions allowed us to evaluate the strength of support for identified orthologues. Neighbor-joining trees based on full-length connexin sequences and connexin sequences lacking the carboxy tail yielded highly similar topologies, both at the level of orthologous gene clades (i.e., the tip nodes) and at the broader level of connexin class (i.e., the internal nodes). A neighbor-joining tree based on the C-terminus (connexin sequences lacking the transmembrane domains and internal loop) was poorly resolved in comparison to the full-length and N-terminal trees. Notably, however, 10 of the 11 orthologous gene clades were identified in the C-terminal tree. The average bootstrap support for these 10 groups was 87.3%. The single incongruity was based on the status of zebrafish Cx27.5. Full-length analysis indicated that Cx27.5 is paralogous to zebrafish Cx31.7 and that these 2 zebrafish connexins are co-orthologues of human Cx32. Analysis of the C-terminus identified zebrafish Cx34.5 (a novel connexin) as the closest relative of Cx27.5; however, this relationship received <50% bootstrap support. Relationships *among* the gene groups, the internal nodes, were largely unresolved, indicating extensive divergence among the C-termini of connexin sequences. However, it is remarkable that the C-termini of zebrafish and mammalian orthologues, separated by at least 400 million years, remain similar, whereas the C-termini of many putative paralogues in both the mammalian and the fish lineages are not identified as close relatives. Close sequence identity among the C-termini of identified orthologues may indicate stronger selection for C-terminus function in orthologues vs. paralogues. Furthermore, duplications may have led to relaxed selection on one of the paralogues, permitting greater divergence among paralogues and facilitating the evolution of novel gene function [13].

An analysis using the optimality criterion of parsimony identified all 11 orthologous gene clades, including the 16 zebrafish orthologues, that were identified in the full-length neighbor-joining analysis (average bootstrap support = 89.6%; Fig. 1). Nodes identifying the 7 additional zebrafish connexins with close human relatives were also well supported in the parsimony tree.

The orthologous relationships we identified are largely consistent with the assignments reported by others [14–18]. We revealed additional orthologues here: zebrafish Cx27.5 is orthologous to human Cx32, and zebrafish Cx55.5 is orthologous to human Cx59, which differs with the findings of one report [14]. Our survey of three complete connexin families permitted the identification of the orthologues as highly supported phylogenetic relationships, rather than case by case assignments based strictly on amino acid sequence identity. This difference may account for the reported disparity in orthology assignments for these connexins.

One other notable difference in orthology assignments is for the zebrafish Cx45 proteins. Previously, zebrafish Cx43.4 was described as most closely related to mammalian Cx45 [34]. At this time, we find that a formerly undescribed sequence, Cx52.8, is the closest orthologue to mammalian Cx45 (Fig. 1). Our phylogenetic analysis revealed four additional zebrafish connexins (plus Cx43.4) whose closest mammalian orthologues are unclear (and therefore termed novel in this report), but are clearly related to the mammalian Cx45/Cx47 group. Therefore, it remains possible that one or more of these zebrafish connexins will share expression and/or functional characteristics of either mammalian Cx45 or Cx47. Indeed, zebrafish Cx43.4 has been shown to exhibit transjunctional voltage properties similar (but not identical) to those of mammalian Cx45 [19], supporting the hypothesis that Cx43.4 is related to, but is not the closest relative of, mammalian Cx45.

Maintenance and loss of human connexins

Four human connexins, *CX37*, *CX31.9*, *CX47*, and *CX31.3*, appear to be absent from the zebrafish genome. These connexins may have been lost from the zebrafish lineage or may have arisen in the mammalian lineage from gene duplication events specific to that group. Alternatively, these genes may be found in the remaining ~4% of the genome sequence that is absent from the current assembly (http://www.sanger.ac.uk/Projects/D_rerio/Zv5_assembly_information.shtml).

Both connexins previously labeled “human-specific” because they are absent from the rat and mouse genomes (i.e., *CX25* and *CX59* [20]), have been identified in the dog, opossum, and cow genome projects, revealing that these genes are not specific to the human genome [21]. We also provide evidence that the zebrafish genome has relatives for *CX25* (i.e., *cx28.8*) and *CX59* (i.e., *cx55.5* and *cx52.9*), further suggesting that each of these genes was present on an ancestral chromosome and that they are not recent additions to the mammalian lineage. In contrast, the single “mouse-specific” connexin, *Cx33* (i.e., absent from the human genome [20]), is also absent from the zebrafish genome, supporting the hypothesis that mouse *Cx33* arose after the divergence of the mammalian and fish lineages.

Evolution of the connexin gene family

Although the zebrafish genome contains almost twice the number of connexins as the mammalian genome, not all human

connexins are found in the zebrafish genome, and others are found in duplicate (or more) copies. Therefore, the large number of zebrafish connexins is not due to a simple whole-genome duplication event followed by loss of a small number of connexins. Rather, the zebrafish has single relatives for some human connexins, multiple relatives for others, and 14 apparently novel connexins, suggesting the occurrence of zebrafish-specific gene duplication events, or the entire loss of these connexin types in the mammalian lineage.

Evidence supporting zebrafish-specific gene duplication events is found both in our phylogenetic tree and in the genome, as fish-specific clades identified in the phylogeny correspond to clusters of physically linked connexins in the zebrafish genome. For example, the connexins in the fish-specific clade containing Cx32.3, Cx31.9, Cx28.9, Cx28.1, and Cx34.5 are physically linked and adjacent to one another, suggesting that they arose by tandem duplications. This cluster is also linked to a sixth, less related connexin, Cx43 [22]. The connexins in this clade are not closely related to Cx43; instead the closest relative is Cx37. However, as this relationship does not receive strong bootstrap support (<50%, Fig. 1), the ancestor for this clade is not clear.

A second fish-specific clade contains four connexins: Cx44.2, Cx45.1, Cx44.6, and Cx43.4. Three of these connexin genes are physically linked and adjacent to one another on the same genomic BAC (*cx44.2*, *cx45.1*, and *cx44.6*), suggesting that the cluster of three also arose by tandem duplications. The three connexins in this cluster are also uniquely similar (90–93% identity), but not identical. One explanation for this similarity is that this region of the genome was "under-assembled" and this cluster of three connexins should overlap (i.e., and represent a single connexin). However, comparison of the flanking regions and introns reveals that the noncoding sequences are related but not identical (82–90% similarity). Still, the apparent divergence of the intronic and flanking noncoding sequence could be due to poor sequence quality. Therefore, we next compared noncoding regions on the finished BAC to the independently derived trace files from the WGS sequencing project. This identified highly related sequences (>98% identical) for the noncoding sequences associated with all three connexins, indicating that the sequence in the BAC is accurate and not the result of the misassembly of poor sequence reads (data not shown). Therefore, these three genes do appear to be the result of a recent series of local duplications. Since the noncoding regions in zebrafish are more similar to one another (i.e., 80–90%) than the average similarity for noncoding sequence observed in comparison of zebrafish to *Fugu* or *Tetraodon*, it is likely that these duplications arose after the divergence of these teleost lineages and, further, that the connexin gene family is continuing to grow in zebrafish.

Connexin gene duplication events are not limited to the zebrafish genome. We also find corresponding phylogenetic and genomic evidence for relatively recent tandem duplication events in the mammalian lineage (see also [23]). The connexins *CX31.1* and *CX30.3*, which form a clade, are adjacent to each other on human chromosome 1 and mouse chromosome 4. Similarly, the connexins *CX26* and *CX30* are adjacent to each other on human chromosome 13 and mouse chromosome 14. The topology of our tree suggests that both of these duplication events occurred after the split between fish and mammals, as we find only a single zebrafish relative for each of these gene pairs. The topology further shows that the duplication likely occurred prior to the split between mouse and human, as both mouse and human contain representatives of these linked genes. Thus, combining phylogenetic and genomic data, we conclude that zebrafish and mammalian connexin genes have undergone multiple independent duplication events. The evolutionary mechanisms regulating connexin number therefore do not appear specific to either lineage but rather represent a more global means of influencing the connexin gene family. It is tempting to speculate that the duplication events that contribute to connexin number also permit the continued specialization of gap junction channels in species- and (or) tissue-specific manners.

Genomic organization of physically linked connexins is maintained

We find four different clusters of zebrafish connexins that are similar to three connexin clusters found in the human genome (Fig. 2), suggesting that the genomic organization of connexins has been maintained throughout evolution. The human genome has a single cluster containing *CX40* and *CX50* on chromosome 1 (1q21.1). The zebrafish genome has two clusters that represent duplicate copies of the *CX40* and *CX50* cluster (Fig. 2A). Note that in one of the zebrafish clusters the orientation of the two genes is inverted (*cx41.8*, *cx44.1*), suggesting an additional chromosomal rearrangement on the zebrafish chromosome. Interestingly, we find additional conservation of synteny for the latter cluster since the *BCL9* gene is found next to *CX40* on human chromosome 1 and the zebrafish orthologue for *bcl9* is found next to zebrafish *cx41.8*. A second human cluster containing four connexins (*CX31.1*, *CX30.3*, *CX31*, and *CX37*) is also found on human chromosome 1 (Hsa1) at 1p35.1. The zebrafish orthologue for *CX31* (i.e., *cx35.4*) is linked to *cx34.4*, a connexin equally related to the human *CX31* neighbors *CX31.1* and *CX30.3* (Fig. 2B). Therefore, the closest zebrafish relative of duplicated human genes is found in the same physical location in the zebrafish genome. Additional evidence supporting this assertion is found by the conservation of synteny in these regions. Two genes, *znf593* and *SEPN1*, are found on Hsa1 upstream of *CX31.1*. Zebrafish orthologues for these same two genes are found local to zebrafish *cx35.4* and *cx34.4* (Fig. 2B). The relative location of these genes suggests at least one rearrangement in this region. Similar to the analysis above, a cluster of three connexins (*CX46*, *CX26*, and *CX30*) is found on human chromosome 13 (Hsa13) at 13q11–q12. The zebrafish orthologue for *CX46* (*cx48.5*) is linked to *cx33.8*, a connexin equally related to the human *CX46* neighbors *CX26* and *CX30* (Fig. 2C). Syntenic analysis also supports these relationships. Human *XPO4* is found adjacent to *CX30* on Hsa13 and the zebrafish orthologue for *xpo4* is found adjacent to zebrafish

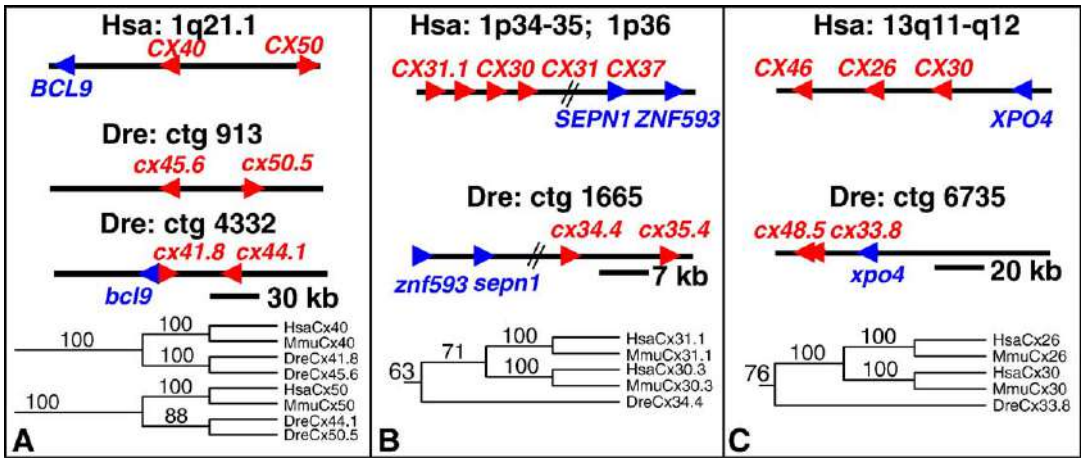

Fig. 2. Conserved genome organization of connexin clusters. (Top) Linked human connexins and linked orthologous zebrafish connexins. Human gene distances are indicated by chromosome position. Zebrafish gene distances are indicated by the number of kilobases. Connexin genes are represented by red arrows and text; non-connexin genes located nearby are represented by blue arrows and text. The (//) on zebrafish contig 1665 represents approximately 150 kb. (Bottom) Phylogenetic relationships for the linked mammalian and zebrafish connexins. Hsa, *Homo sapiens*; Mmu, *Mus musculus*; Dre, *Danio rerio*. Ctg, contig.

cx33.8. These data therefore provide strong evidence that each cluster of the zebrafish connexins and their respective related human connexins (i.e., zebrafish *cx34.4* and human *CX31.1* and *CX30.3*; or zebrafish *cx33.8* and human *CX26* and *CX30*) evolved from a single common ancestor.

Phylogenetic clades represent connexin classes

Human connexin genes are named after their assignment into α, β, or γ classes. However, the use of various criteria to identify these classes has resulted in ambiguous assignments for some connexins. Originally, the assignment of connexin sequences to classes was based on either a K-X-X-X-E motif (α) or an R-X-X-X-E motif (β) lining the predicted channel forming region of the connexin proteins [24]. Subsequent criteria included overall similarity to previously classified connexins in combination with the overall length of the predicted polypeptide (i.e., since α connexins tended to be longer than β connexins [25]). However, some connexins exhibit the channel motif for one class and overall similarity to a different class [3], suggesting that neither of these methods reflect an appropriate means to distinguish classes. Still, there is little doubt that connexins fall into classes and that connexins in the same classes tend to cluster together on phylogenetic trees [3,26–29]. Indeed, Bennett et al. [26] described two groups of connexins based on their distinction in a phylogenetic tree (where group I represented the β connexins and group II represented the α connexins). Others have used trees to facilitate class assignments when strict sequence comparisons were ambiguous [28,29].

Our findings further support the use of phylogenetic analysis to reveal evolutionarily relevant groups of connexins. Connexin genes previously identified as belonging to the α, β, or γ classes were largely contained within distinct clades (colored shading in Fig. 1). The clade containing the β class was the most highly supported (Fig. 1, green shading). All human and mouse connexins previously identified as belonging to the β class, and their zebrafish orthologues, were found in a single, highly supported clade. The γ clade identified in our analysis is also largely consistent with previous reports (Fig. 1, dark blue shading), with one exception. Human Cx31.3 and its mouse orthologue Cx29 grouped with the γ connexins in our analysis, albeit weakly (bootstrap support 63%). The clade representing the α class (Fig. 1, pink shading) contains most of the human and mouse connexins, and their zebrafish orthologues, classified as α's. One exception is human Cx59 (gene name *GJA10*), which grouped in a separate clade with human Cx62/mouse Cx57 (see below). A second exception is human Cx31.9/mouse Cx30.2, which has been classified both as an α connexin (i.e., *GJA11* [30]) and as an "ungrouped" connexin [29]. Its current gene name, *GJC1*, signifies that its classification is somewhat tenuous. We did not find support for the inclusion of these connexins (Cx31.9/Cx30.2) with either the α or the γ clades.

We identified a formerly unclassified grouping in our analyses (Fig. 1, yellow shading). This clade contains human Cx62 and Cx59, as well as mouse Cx57 and four zebrafish connexins: Cx52.6 and Cx52.7 (orthologues of Cx62), and Cx52.9 and Cx55.5 (orthologues of Cx59). This clade was highly supported (average bootstrap support 97%) and thus represents a potential fourth class of connexin genes. Indeed, the future addition of newly identified connexin sequences from more species may demonstrate that connexins not found in these four larger clades represent the first members of undetermined groups.

High statistical correspondence between clades and connexin classes indicates that rigorous phylogenetic analysis provides the best means of identifying evolutionarily real, and likely functionally significant, connexin classes. As mentioned above, similar clades have been identified previously for the human and mouse connexins [23,26]; however, these clades have not been used as a criterion to distinguish the classes. The inclusion of a third complete connexin gene family from a distantly related vertebrate lineage both validates the use of clades to specify connexin classes and demonstrates that the grouping of connexins is a general feature of connexin gene families across all vertebrates.

Conclusion

This is the first report that compares entire connexin families from three species, permitting a broad comparative analysis of 76 connexins from human, mouse, and zebrafish. Phylogenetic analysis revealed robust orthologous relationships from zebrafish to human, provided evidence for local duplication events in each genome leading to the growth of the connexin family, and suggested that deeper clades in phylogenetic trees represent the separation of the connexin classes. We further infer general mechanisms guiding the evolution of the connexin gene family, including mechanisms for gene retention, loss, and expansion. Indeed, continuing tandem duplication events of connexin genes may lead to adjustments in gap junction composition in one or more tissues, resulting in increasingly more specialized gap junctional communication.

Materials and methods

Genome search for new connexin sequences

Nucleotide sequences of the 16 reported zebrafish connexin genes (Table 1) were compared (using BLASTN) against five available databases in the following order: the whole-genome shotgun assembly, version 5 (WGS, v5) via Ensembl, the finished and unfinished genomic BAC sequences, the trace files associated with the WGS project, the zebrafish EST database, and the NCBI gene database.

For all connexin sequences identifying one or more sequenced genomic BACs (i.e., with 60–100% identity to the query connexins), we used intron/exon prediction software (GENSCAN; http://genes.mit.edu/GENSCAN.html and FGENESH; http://sun1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind) to identify predicted coding regions across the entire BAC. All of the predicted sequences were examined using the conserved domain (CD) function via NCBI [31] to identify proteins containing connexin domains.

Derivation of connexin genes from available sequences

Our searches yielded 37 putative connexin sequences (see Supplementary Material for predicted coding and peptide sequences for all 37 connexins). Full-length coding sequences from the 16 reported connexins were identified using published accession numbers (Table 1). Full-length coding sequences for the remaining connexins were derived by evaluating the sequences from the relevant databases for each gene. Eight connexin sequences were found in the NCBI gene database as full-length mRNA sequences; corresponding protein sequences were utilized for this group (*cx52.9*, *cx28.6*, *cx30.9*, *cx33.8*, *cx34.4*, *cx35.4*, *cx44.2*, and *cx47.1*). Eight genes were identified both from the sequenced BACs and from the WGS sequence as partial or full-length predicted transcripts (*cx39.4*, *cx41.8*, *cx50.5*, *cx44.6*, *cx45.1*, *cx52.8*, *cx34.1*, and *cx40.5*). The entire protein coding sequence for this group was derived from the appropriate BAC using corroborating GENSCAN and FGENESH predicted transcripts. Full-length predicted transcripts were identified similarly for the three sequences found only from the sequenced BACs (i.e., *cx31.7*, *cx46.8*, and *cx28.8*). The final two connexins were only found in the WGS sequence: one as a full-length predicted transcript (*cx52.7*) and one as a partial predicted transcript (*cx35.8*). The partial predicted transcript for *cx35.8* was evaluated as described below to determine whether or not to include its sequence in further analyses.

Alignment of connexin sequences

All full-length amino acid sequences for the human, mouse, and zebrafish connexins were aligned using ClustalW (http://www.ebi.ac.uk/clustalw/, see Supplementary Material). Predicted transmembrane-spanning domains were labeled, revealing the two extracellular loops (including the three conserved cysteine residues), the intracellular loop, and the carboxy-terminus. Closer examination of this alignment of 76 connexins revealed two zebrafish connexins (Cx39.4 and Cx35.8) that required further scrutiny.

Cx39.4 begins with the sequence "MSRADWG," where R and A represent insertions specific to Cx39.4. Sequence from the finished BAC zC261O1, the single predicted transcript from the WGS project and six overlapping trace files all predict the same amino acid sequence, suggesting that the additional two amino acids are not the result of sequencing error but instead may represent a new feature for this connexin.

Alignment of the two overlapping predicted transcripts for gene Cx35.8 (ENSDART00000016465, GENSCAN00000015084) revealed that approximately six amino acids are missing from the amino terminus (the presence of a stop codon indicated that the remaining sequence was complete). Additional trace files or ESTs did not extend this sequence. Since it is clear that only a small number of amino acids are missing from the predicted peptide, the truncated sequence was included in the phylogenetic analysis.

Phylogenetic analysis

To determine the phylogenetic relationships between putative zebrafish connexins and mammalian connexins, nucleotide coding regions and protein sequences for zebrafish, mouse, and human connexins were independently aligned in ClustalW. A distance matrix was generated from the amino acid sequences using standard mean differences in the phylogenetic program PAUP* 4.0 [32]. A neighbor-joining phylogram was generated from the distance matrix, using human pannexin1 (NP_056183) as the outgroup. The amino acid sequences of the C-terminus and transmembrane domain region (including intracellular loop) were independently aligned and similarly analyzed. A parsimony reconstruction based on amino acid sequence was also generated for the full alignment using a modified version of the PROTPARS executable in PHYLIP [33], where gaps were treated as missing data. A heuristic search based on 10 random addition sequences was conducted; TBR branch swapping was in effect. Bootstrap values for the distance tree are based on 1000 neighboring replicates. Bootstrap values for the parsimony tree are based on 1000 replicates using "fast" stepwise addition.

Identification of linked connexins in the zebrafish genome

All BACs containing connexin genes were located on the fingerprinted clone map (http://www.sanger.ac.uk/cgi-bin/Projects/D_rerio/WebFPCreport.cgi) to identify neighboring BACs within contigs. Connexins identified on the same BAC (and at different nucleotide positions) or on different BACs that locate to the same contig are physically linked. Linked human connexins were described in Willecke et al. [20] and nucleotide locations were identified using NCBI MapViewer (http://www.ncbi.nlm.nih.gov/mapview/).

Syntenic analyses

To find evidence for the conservation of synteny, we compared genes neighboring the zebrafish connexins (i.e., the 14 novel zebrafish connexins and the 7 zebrafish connexins closely related to mammalian connexins) to the genes neighboring the human connexins. Local genes were identified for each zebrafish connexin by performing GENSCAN analysis on entire BACs containing each connexin (18 of these genes were located on BACs). For the 3 connexins not located to BACs, genes within 200 kb of the connexin were identified on the zebrafish genome assembly, Zv5. Putative orthologues for each zebrafish gene were located on the human map using the human genome browser at UCSC (http://genome.ucsc.edu/). Positions were compared to the location for each human connexin [4] and connexins located nearby were noted.

Acknowledgments

The authors are especially appreciative to Todd Oakley for advice regarding parsimony analyses of amino acid sequences. The authors thank Alex Brands and members of the Iovine lab

for critically reading and discussing the manuscript, and the Zebrafish Nomenclature Committee for their advice regarding zebrafish connexin nomenclature. This work was supported by the NIDCR (5K22DE014863 to M.K.I.), the NIGMS (GM56988 to T.H.P.C. and GM55725 to M.M.F.), and the Bioengineering and Bioscience 2020 Funds (M.K.I., M.M.F., T.C.M.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2005.10.005.

References

[1] L. Makowski, et al., Gap junction structures: II. Analysis of the X-ray diffraction data, J. Cell Biol. 74 (2) (1977) 629–645.
[2] D.A. Gerido, T.W. White, Connexin disorders of the ear, skin, and lens, Biochim. Biophys. Acta 1662 (1/2) (2004) 159–170.
[3] N.M. Kumar, N.B. Gilula, Molecular biology and genetics of gap junction channels, Semin. Cell Biol. 3 (1) (1992) 3–16.
[4] G. Sohl, K. Willecke, Gap junctions and the connexin protein family, Cardiovasc. Res. 62 (2) (2004) 228–232.
[5] C. Elfgang, et al., Specific permeability and selective formation of gap junction channels in connexin-transfected HeLa cells, J. Cell Biol. 129 (3) (1995) 805–817.
[6] C.G. Bevans, et al., Isoform composition of connexin channels determines selectivity among second messengers and uncharged molecules, J. Biol. Chem. 273 (5) (1998) 2808–2816.
[7] F.F. Bukauskas, et al., Coupling asymmetry of heterotypic connexin 45/connexin 43–EGFP gap junctions: properties of fast and slow gating mechanisms, Proc. Natl. Acad. Sci. USA 99 (10) (2002) 7113–7118.
[8] A. Amores, et al., Zebrafish hox clusters and vertebrate genome evolution, Science 282 (5394) (1998) 1711–1714.
[9] R.L. Gimlich, N.M. Kumar, N.B. Gilula, Differential regulation of the levels of three gap junction mRNAs in *Xenopus* embryos, J. Cell Biol. 110 (3) (1990) 597–605.
[10] B. Risek, et al., Modulation of gap junction transcript and protein expression during pregnancy in the rat, J. Cell Biol. 110 (2) (1990) 269–282.
[11] E.C. Beyer, D.L. Paul, D.A. Goodenough, Connexin43: a protein from rat heart homologous to a gap junction protein from liver, J. Cell Biol. 105 (6, Pt. 1) (1987) 2621–2629.
[12] S.A. Farber, et al., The zebrafish annexin gene family, Genome Res. 13 (6A) (2003) 1082–1096.
[13] S. Ohno, Evolution by Gene Duplication, Springer Verlag, New York, 1970.
[14] R. Dermietzel, et al., Molecular and functional diversity of neural connexins in the retina, J. Neurosci. 20 (22) (2000) 8331–8343.
[15] E. McLachlan, et al., Zebrafish Cx35: cloning and characterization of a gap junction gene highly expressed in the retina, J. Neurosci. Res. 73 (6) (2003) 753–764.
[16] S. Cheng, T. Christie, G. Valdimarsson, Expression of connexin48.5, connexin44.1, and connexin43 during zebrafish (*Danio rerio*) lens development, Dev. Dyn. 228 (4) (2003) 709–715.
[17] T.L. Christie, et al., Molecular cloning, functional analysis, and RNA expression analysis of connexin45.6: a zebrafish cardiovascular connexin, Am. J. Physiol. Heart Circ. Physiol. 286 (5) (2004) H1623–H1632.
[18] G. Zoidl, et al., Molecular cloning and functional expression of zfCx52.6: a novel connexin with hemichannel-forming properties expressed in horizontal cells of the zebrafish retina, J. Biol. Chem. 279 (4) (2004) 2913–2921.
[19] T. Desplantez, et al., Characterization of zebrafish Cx43.4 connexin and its channels, Exp. Physiol. 88 (6) (2003) 681–690.
[20] K. Willecke, et al., Structural and functional diversity of connexin genes in the mouse and human genome, Biol. Chem. 383 (5) (2002) 725–737.
[21] V. Cruciani, S.O. Mikalsen, The connexin gene family in mammals, Biol. Chem. 386 (4) (2005) 325–332.
[22] M.K. Iovine, et al., Mutations in connexin43 (GJA1) perturb bone growth in zebrafish fins, Dev. Biol. 278 (1) (2005) 208–219.
[23] M.V. Bennett, X. Zheng, M.L. Sogin, The connexins and their family tree, Soc. Gen. Physiol. Ser. 49 (1994) 223–233.
[24] L.C. Milks, et al., Topology of the 32-kd liver gap junction protein determined by site-directed antibody localizations, EMBO J. 7 (10) (1988) 2967–2975.
[25] J. Eiberger, et al., Connexin genes in the mouse and human genome, Cell Commun. Adhes. 8 (4–6) (2001) 163–165.
[26] M.V. Bennett, et al., Gap junctions: new tools, new answers, new questions, Neuron 6 (3) (1991) 305–320.
[27] G. Sohl, et al., The murine gap junction gene connexin36 is highly expressed in mouse retina and regulated during brain development, FEBS Lett. 428 (1/2) (1998) 27–31.
[28] J. O'Brien, et al., Cloning and expression of two related connexins from the perch retina define a distinct subgroup of the connexin family, J. Neurosci. 18/19 (1998) 7625–7637.
[29] T.W. White, et al., Virtual cloning, functional expression, and gating analysis of human connexin31.9, Am. J. Physiol. Cell Physiol. 283 (3) (2002) C960–C970.
[30] P.A. Nielsen, et al., Molecular cloning, functional expression, and tissue distribution of a novel human gap junction-forming protein, connexin-31.9. Interaction with zona occludens protein-1, J. Biol. Chem. 277 (41) (2002) 38272–38283.
[31] A. Marchler-Bauer, S.H. Bryant, CD-Search: protein domain annotations on the fly, Nucleic Acids Res. 32 (2004) W327–W331 (Web Server issue).
[32] D.L. Swofford, PAUP*, Phylogenetic Analysis Using Parsimony (*and other methods), Version 4, Sinauer Associates, Sunderland, MA, 2002.
[33] J. Felsenstein, PHYLIP Phylogeny Inference Package, Cladistics 5 (1989) 164–166.
[34] J.J. Essner, et al., Expression of zebrafish connexin43.4 in the notochord and tail bud of wild-type and mutant no tail embryos, Dev. Biol. 177 (2) (1996) 449–462.
[35] M.U. Hussain, et al., Transcriptional and translational regulation of zebrafish connexin 55.5 (zf.Cx.55.5) and connexin 52.6 (zf.Cx52.6), Cell. Commun. Adhes. 10 (4-6) (2003) 227–231.
[36] B. Chatterjee, et al., Developmental regulation and expression of the zebrafish connexin43 gene, Dev. Dyn. 233 (3) (2005) 890–906.
[37] N. Cason, et al., Molecular cloning, expression analysis, and functional characterization of connexin44.1: a zebrafish lens gap junction protein, Dev. Dyn. 221 (2) (2001) 238–247.
[38] S. Cheng, et al., Connexin48.5 is required for normal cardiovascular function and lens development in zebrafish embryos, J. Biol. Chem. (2004).
[39] V. Valiunas, et al., Biophysical characterization of zebrafish connexin35 hemichannels, Am. J. Physiol. Cell Physiol. 287 (6) (2004) C1596–C1604.