Genomic annotation and transcriptome analysis of the zebrafish (Danio rerio) hox complex with description of a novel member

Corredor-Adamez, M.; Welten, H.P.; Spaink, H.P.; Jeffery, J.E.; Schoon, R.T.; Bakker, M.A.G.; ... ; Richardson, M.K.

EVOLUTION & DEVELOPMENT 7:5, 362–375 (2005)

Genomic annotation and transcriptome analysis of the zebrafish (*Danio rerio*) *hox* complex with description of a novel member, *hoxb13a*

M. Corredor-Adámez, M. C. M. Welten, H. P. Spaink, J. E. Jeffery, R. T. Schoon, M. A. G. de Bakker, C. P. Bagowski, A. H. Meijer, F. J. Verbeek, and M. K. Richardson*

Institute of Biology, Leiden University, Wassenaarseweg 64, 2333 AL Leiden, The Netherlands

*Author for correspondence (email: richardson@rulsfb.leidenuniv.nl)

SUMMARY The zebrafish (*Danio rerio*) is an important model in evolutionary developmental biology, and its study is being revolutionized by the zebrafish genome project. Sequencing is at an advanced stage, but annotation is largely the result of in silico analyses. We have performed genomic annotation, comparative genomics, and transcriptional analysis using microarrays of the *hox* homeobox-containing transcription factors. These genes have important roles in specifying the body plan. Candidate sequences were located in version Zv4 of the Ensembl genome database by TBLASTN searching with *Danio* and other vertebrate published Hox protein sequences. Homologies were confirmed by alignment with reference sequences, and by the relative position of genes along each cluster. RT-PCR using adult Tübingen cDNA was used to confirm annotations, to check the genomic sequence and to confirm expression in vivo. Our RT-PCR and microarray data show that all 49 *hox* genes are expressed in adult zebrafish. Significant expression for all known *hox* genes could be detected in our microarray analysis. We also find significant expression of *hox8* paralogs and *hoxb7a* in the anti-sense direction. A novel gene, *D. rerio hoxb13a*, was identified, and a preliminary characterization by in situ hybridization showed expression at 24 hpf at the tip of the developing tail. We are currently characterizing this gene at the functional level. We argue that the oligo design for microarrays can be greatly enhanced by the availability of genomic sequences.

INTRODUCTION

The zebrafish *Danio rerio* is an important model system for biomedical research. Furthermore, its short life cycle, small size, transparent embryos, and high fecundity makes it a suitable and easily maintained species for studies in developmental biology. Genes can be knocked down with anti-sense morpholinos, and a wide range of mutants is available. Zebrafish resources have recently been strengthened by the zebrafish genome project, which is based on DNA from the *Tübingen* zebrafish strain.

Comparative studies of the zebrafish genome are allowing putative regulatory sequences to be located (Hadrys et al. 2004), and may provide insights into evolutionary processes and events, including the fin/limb transition (Sordino et al. 1995; Duboule and Sordino 1996). Like other teleosts, the zebrafish shows numerous gene duplications relative to other vertebrates. This is probably because of a whole-genome duplication event in ancestral ray-fin fishes (Amores et al. 1998; Meyer and Schartl 1999; Aparicio 2000; Malaga-Trillo and Meyer 2001; Robinson-Rechavi et al. 2001; Taylor et al. 2001; Venkatesh 2003; Hoegg et al. 2004).

According to classical models of gene evolution (Ohno 1970), most of the duplicated genes are destined to be lost rapidly in the aftermath of the duplication event. One copy of the gene may retain the original function, whereas the other becomes a pseudogene or is lost altogether. However, in some cases the duplicated genes are retained and the duplication–degeneration–complementation (DDC) model (Force et al. 1999) has been suggested to explain the selective maintenance of duplicated genes. Because of the advantages offered by zebrafish for genetic research, the DDC model has been experimentally validated (e.g., Bruce et al. 2001; McClintock et al. 2002).

One group of genes that is of particular interest in developmental genetics and evolution are the *hox* genes. These are homeobox-containing transcription factors, first discovered in *Drosophila melanogaster* (Lewis 1978). Mutations in these genes cause homeotic mutations affecting segment identity (Bateson 1894, reprinted in 1992).

An intriguing feature of this gene family is its close genetic linkage forming the *hox* clusters or homeotic complexes. The relative position of genes in the cluster correlates with their spatial pattern of expression along the anteroposterior axis of the embryo (Lewis 1978; Akam 1987). This spatial colinearity—where the 3′ end of the cluster corresponds to rostral expression along the primary axis of the embryo—also corresponds to a temporally colinear pattern of expression, such

that anterior genes are expressed earlier in development. There is evidence that colinearity is related to competition between genes for a remote enhancer that preferentially recognizes 5′ members (Kmita et al. 2002).

Temporal and spatial patterns of expression could be linked. If so, then heterochrony in *hox* gene expression could directly lead to a change in their spatial pattern and therefore affect pattern formation and morphogenesis (Duboule 1994; van der Hoeven et al. 1996). However, in the tunicate *Oikopleura dioica*, spatial colinearity is probably present, even though the *hox* genes are not clustered (Seo et al. 2004).

The importance of *hox* genes in development (including patterning of the primary axis, rhombomeric neural crest, and fin or limb formation; see Duboule 1994) has led to their extensive study. The *hox* cluster is found in all metazoa (de Rosa et al. 1999), and the tandem duplications that gave rise to the complex have been traced. It is likely that the genomic complexity of the *hox* genes is related to morphological complexity (Garcia-Fernandez and Holland 1994; Holland and Garcia-Fernandez 1996; Wagner et al. 2003). Thus, tetrapods have four *hox* clusters (A, B, C, and D), believed to have arisen by serial genome duplications, and this could be linked to the greater morphological complexity of vertebrates compared with metazoa, which have only one cluster.

There was a further genome duplication event in the ray-finned fishes (Amores et al. 1998; Meyer and Schartl 1999; Malaga-Trillo and Meyer 2001; Taylor et al. 2001), and some authors have suggested that morphological diversification in teleosts might have been facilitated by the increased number of *hox* genes (Malaga-Trillo and Meyer 2001; Amores et al. 2004; Hoegg et al. 2004).

The high conservation of the sequence of individual *hox* genes, and their clustered organization, makes them ideal markers for testing alternative scenarios of vertebrate genome evolution (Amores et al. 1998, 2004; Aparicio 2000). The *hox* cluster in ancestral chordates is believed to be formed by 13 or 14 genes—depending on whether the recently discovered vertebrate *hox*14 genes (Powers and Amemiya 2004) are orthologous to amphioxus *hox14* (Ferrier et al. 2000). Subsequent duplication of the whole cluster in the vertebrate lineage led to the formation of the corresponding 13 or 14 paralogous groups. The loss of individual genes or even whole clusters later on has given rise to the current conformations.

Gene mapping data have been gathered from a wide range of vertebrate species to clarify the picture of vertebrate genomic evolution. However, studies of the kind needed to provide the required genetic linkage information are more difficult to perform in nonmodel species. As the ultimate goal is to achieve a physical map of the clusters, genome projects will help to determine the genomic organization of the clusters. The availability of a reliable sequence of all the clusters and adjacent regions provides the material needed for comparative studies aiming to identify and further characterize conserved regulatory regions (Santini et al. 2003; Nobrega and Pennacchio 2004).

In view of the importance of the zebrafish for research in biomedicine, development, and evolution, we acknowledge that it will be important to have a well-characterized picture of its genome, especially of those regions already known to be important in developmental regulation, such as the *hox* clusters. Furthermore, the annotation of these clusters, and the comparison with previously known information, will help fill gaps in our knowledge of the *hox* clusters. Conversely, we can use existing knowledge of the *hox* genes to examine the quality of the reported genomic sequence.

In this article, we describe the genomic organization of known *hox* genes in zebrafish and perform a comprehensive search for unrecorded ones. In addition, we have conducted an initial survey of the existing microarray platforms. We also provide evidence for the existence of a zebrafish member of the paralog 13 group in the *ba* cluster.

MATERIALS AND METHODS

Genomics and sequence analysis

We used the TBLASTN algorithm to identify potential *hox* genes in the most recent zebrafish genome build (Zv4). These sequence data were produced by the Zebrafish Sequencing Group at the Sanger Institute and can be obtained from http://www.ensembl.org/Danio_rerio/. The query sequences were a variety of zebrafish *hox* reference proteins (Table 1). The homology of candidate *hox* genomic sequences was examined by comparison with a variety of reference proteins, from different vertebrate species, using the AlignX algorithm. Open reading frames (ORFs) were predicted from several lines of evidence, including Genscan predictions and alignment with reference proteins from zebrafish and other vertebrates. The alignment in Fig. 1 was produced using Clustal W v. 1.83 (http://www.ebi.ac.uk/clustalw/index.html).

Further validation of genomic annotations was made by comparing coding sequences (DNA) or exons 1 and 2 (translations) from *D. rerio* (Zv4) with those from *Homo sapiens* (NCBI genome build 34 version 3). We used two search methods: parsimony, using the parsimony ratchet in PAUP 4 (Swofford 1999), and likelihood, using quartet puzzling in TreePuzzel 5.2. (Schmidt et al. 2002; Strimmer and von Haeseler 1996). DNA sequences were aligned using their predicted amino acid sequences, according to the method of Bininda-Emonds (2005).

Sequenced PCR products were used to refine the ORF predictions, and to confirm that the genes were expressed in vivo. Where the genomic sequence and reference proteins were at variance, the results of PCR were used to make a final decision.

RT-PCR

Total RNA was isolated from adult *Tübingen* zebrafish with Trizol (Gibco/Invitrogen, Breda, The Netherlands) according to the manufacturers' protocol, followed by DNase I treatment (Invitrogen, Breda, the Netherlands). Reverse transcriptase reactions were per-

Table 1. Evidence used to make the annotations reported in this article

Zfin gene	Ref. proteins	ESTs	PCR	
			Forward primer	Reverse primer
hoxa1a	NP_571611.1	NM_131536	CGTCATGGTTACAGTAGCGGAAAC	GCTGCCCACCAAATCCATACT
hoxa3a	NP_571609.1	NM_131534	AATGTCCGAGGTCTGCCAACA	TTGTCTCCAGCGCAGCTCTCT
hoxa4a	AAD15939.1	DRHOXX4	TGTCAACCAACGATCTCAACGA	TCTTGGGCACTCCTCCAGAGT
hoxa5a	NP_571615.1	NM_131540	ATCCTTAGCCAACTCTCCTGTCA	TTTCCCCTCCGGTCCGGCCAA
hoxa9a	NP_571607.1	NM_131532	TCTCCACTGGTGTACCGTCAGT	GCATGAAGCCAGTTGGACACA
hoxa10a		DREHOXB10		
hoxa11a	NP_571619.1	NM_131544	GGAGGCTCGGACAGTGTAGTAGA	GGGACACCTCTTCTTACGAAACC
hoxa13a	AAQ72839.1			
hoxa2b	NP_571181.1	NM_131106	TCCTTTTGAGCAGACCATTCCA	CGCCGTGATGATTCTCTTTGC
hoxa9b	NP_571608.1	NM_131533	CAGGAGACAGCGATGGAGCAT	GCCAGTTGGACGAAGGGTTA
hoxa10b	CAD59110.1	DREHOXA10	TGTCTCAGACATGCACCACATCT	GCCGTCAGCCAGTTAGCTGTAC
hoxa11b	NP_571222.1	NM_131147	TACGGCAGCGTGGGGAGGAA	CCTTTTTTTCCGTGTCTTTTGTC
hoxa13b	NP_571269.1	NM_131194	ND	
hoxb1a	NP_571190.1	NM_131115	ND	
hoxb2a	NP_571191.1	NM_131116	GAAGGAGAAAAAATCCTCGAAGAAG	CAAACTACACACCACCGGGAC
hoxb3a	NP_571192.1	DRE537509	Incomplete genomic sequence	
hoxb4a	NP_571193.1	NM_131118	ND	
hoxb5a	NP_571176.2	NM_131101	TTCGCCCCCTTCCGACCAAAG	TTTTGCCGTCTGGTCCAGTCA
hoxb6a	NP_571194.1	NM_131119	ACCCATCCTCCTTTTACCAACA	GACCCCTTCGACCAGCGTTAC
hoxb7a	CAD59112.1	DREHOXB7	GCACCGGTCTCTTCATCATCTTC	TGGTAGCGGGAATAGGTCTGA
hoxb8a	NP_571195.1	NM_131120	GCGGATTTGCTCAGGACCTA	CTGATAGCGGCTGTAGGTCTGA
hoxb9a	NP_571196.1	NM_131121	ND	
hoxb10a	AAD15943.1	DRHOXB1	Incomplete genomic sequence	
hoxb13a	NONE		ATACCCCGGTCACGTCTGAAG	GAATGAGCCCCCGTCATGTTG
hoxb1b	NP_571217.1	NM_131142	TCACTCAAGCAGATGACCACATG	TTCCGTATTCGGCGACTTTAAC
hoxb5b	NP_571612.2	NM_131537	CCAGCGATACTCTGGAATTGAAG	CGTTTTCCATCAGGTCCAGTCA
hoxb6b	CAD44457.1	AL645798.2	GGAGCTACCAATGTCCAAGACAAG	TTCAAGAGTCTGAAACCGAGTGTAG
hoxb8b	CAD44456.1	DREHOXA8 AF071255	TTTCAGCACGCGGCTCAGTTC	TGTAGGTCTGCCTGCCCCTTCT
hoxc1a	NP_571606.1	NM_131531	CCAAGTCCCCTCCTTCCAAAG	TCCGAGCATTTTCCCAGTTGT
hoxc3a	NP_571257.1	NM_131182	GGAAACGACATGCTGAAGAAAG	CCGTACAACTGCTTCACCATTG
hoxc4a	NP_571197.1	NM_131122	ATCCCTGAGCCTGATACTCAAAG	GGGTTCCGCTCCATTGTAACTAGA
hoxc5a	NP_571219.1	NM_131144	AGCAGTTCAGCGAACACAGTCT	TACCGGGTGTAACTGGTTCGT
hoxc6a	NP_571198.1	NM_131123	TTTCGTCTTATGGCACTACAGTGA	ATCTGGCGACCTCTTCTTCTG
hoxc8a	CAA74879.1	BQ826563 DREHOXC8	TCACGTACAGGACTTTTTCCATCA	GGCTGTATGTCTGCCTTCCATTG
hoxc9a	NP_571603.1	NM_131528	CCATCCATACACTCACCAACCTC	CACCCTTGCTACCTCATATCGC
hoxc10a	AAD15950.1	AF071257 DREHOXC10	CCAGGTGTACGCCAGTGAAAG	CCGTTGCTTTTTCCAATTTAGAC
hoxc11a	AAD15951.1	Y14541	CCGTCTCTTCGTTCCTTCCA	TACAGCAACGGATTTCCAGAGA
hoxc12a			GCAGCACCAGAGATTCGTGTTC	TGGAATAGGGTTTTCGTTTCTTG
hoxc13a	NP_571618.1	NM_131543	ACTTGCAGCAGAAACCATGTTC	CACGTCGATAACTGCTGACTTCTG
hoxc6b	NP_571605.1	NM_131530	TCTACCCAGCGTTGCCATCA	GCGAATAGATTTGGCGACCTT
hoxc11b	AAD15952.1	AF071258.1	GAGCATCGGAAAGACCAACGTT	GGACATCGCTTCTTCCTCATTC
hoxc12b	NP_571620.1	AF071260 NM_131545	GGCGGTGGTGACAATAACAGT	TTTCGAGTTTGTCTGTGCATTG
hoxc13b	NP_571621.1	NM_131546	CGGGGACTTCACTGGGCTAT	CGTTCATCTCGGCTTGGTGT
hoxd3a	CAA74286.1	DRHOXD3	CAGGGCAACAGCCAGCCTGAGA	GCGGACTCTTGTCATCGCAGGT
hoxd4a	NP_570113.1	NM_130757	TGTAGCACTGTCCAGGGCTCGT	CAGGTCCTGTGTAATCCGGGTT
hoxd9a	NP_571201.2	NM_131126	GGACACTATTATGGGGCACGA	GGATCCAGTTGGCAGCAGGGTT
hoxd10a	NP_571241.1	NM_131166	GTCCTTTCCCAACAGCTCTCC	ACAGGAGAACATAGATGGACCGA
hoxd11a	CAA61030.1	DREHOXD11	GCCATCACGCTGGTACTAACTCCT	ACGGGCATCTCTTCTTTCTGGACT
hoxd12a	CAA61032.1	CB359704	CTCTCAGCCGTTTTTCAGCAAT	CTGGGGCTTCGTGTAAGGTTT
hoxd13a	NP_571244.2	NM_131169.2	ACGGAGGAGGACTGGATGAAG	AACCCGCTTCTTTCTCCCGC

Note: Zfin gene names can be found at http://zfin.org/. Numerous *hox* reference sequences from other vertebrates were also used to support the annotations, but because of limitations of space only the evidence derived from *Danio rerio* is shown. The GENBANK accession numbers of our PCR products are given in Table 2.
ND, not determined.

```
Spheroides      MTTSLVLNPRWPADTVMFVYENNLDDLNKNMEGLVGSGNFAASQCRNIMAHSAAAAALGG 60
Tetraodon       MTTSLVLNPRWPADTVMFVYENNLDDLNKNMEGLVGSGNFAASQCRNIMAHSAAAAALGG 60
Fugu            MTTSLVLNPRWPADTVMFVYENNLDDLNKNMEGLVGSGNFAASQCRNIMAHSAAAAALGG 60
Danio           MTTSLVLNPRW-ADTVMFVYENNLDEL-KNMEGLVGSGNFAANQCRNLMAHS----ALSG 54
                ********** *************:* **************.****:****    **.*

Spheroides      HPSGLVHSSAGYSTVDVTATSSNETLTSSGKQCVSGPCPGATVPHQSSSAATALPYSYFG 120
Tetraodon       HPSGLVHSSAGYSAVDVAATSSNETLTSSGKQCASGPCPAATVPHQSSSAAAALPYSYFG 120
Fugu            HPSGLVHSSTGYSTVDVTATSSNDTLTSSGKQCATGPCPAATVPHQSSSAATALPYSYFG 120
Danio           HPSSLVHGSS-YPTVDVSTSSSAE----SGKQCT--PCP--TVPQASSTGP--IPYGYFG 103
                ***.***.*: *.:***:::** :    *****.  ***  ***: **:..  :**.***

Spheroides      NGYYPCRMGRGSLKSCTQAAGAALSSQ-YMDTTVNSDDYSNHRAKEFAFYHSYPSPYQSM 179
Tetraodon       NGYYPCRMGRGSLKSCTQAAGAALSSQ-YMDTTVSSDDYSNHRAKEFAFYHSYPSPYQSM 179
Fugu            NGYYPCRMGRGSLKSCTQAAGAALSSQ-YMDTTVNSDEYSNHRAKEFAFYHSYPSPYQSM 179
Danio           NSYYPCIMGRGSLKSCTQPSALSYTAEKYMDTPVTSEEYP-TRAKEFAFYHSYPSPYQSM 162
                *.**** **********.:. : ::: ****.*.*::*.  ******************

Spheroides      ASYLDVSVVQTLGAGEPRHDTLLPMDSYQPWALTNGWGGQMYCSKDQGQAGHLWKSALAD 239
Tetraodon       ASYLDVSVVQTLGAGEPRHDTLLPMDSYQPWALTNGWGGQMYCSKDQGQAGHLWKSALAD 239
Fugu            ASYLDVSVVQTLGAGEPRHDTLLPMDSYQPWALTNGWGGQMYCSKDQGQAGHLWKSALAD 239
Danio           ASYLDVSVVQTLGTGEPRHDSLLPMDSYQPWALANGWGSQMYCSKDQGQAGHLWKSALAD 222
                *************:******:************:****.********************

Spheroides      VVAHQHDGSPFRRGRKKRIPYTKIQLKELEKEYAANKFITKDKRRKISAATNLSERQITI 299
Tetraodon       VVAHQHDGSPFRRGRKKRIPYTKVQLKELEKEYAANKFITKDKRRKISAATNLSERQITI 299
Fugu            VVAHQHDGSPFRRGRKKRIPYTKVQLKELEKEYAANKFITKDKRRKISAATNLSERQITI 299
Danio           VVAHQHDGGSFRRGRKKRIPYTKVQLKELEKEYAANKFITKDKRRKISAVTNLSERQITI 282
                ********..*************:************************.**********

Spheroides      WFQNRRVKEKKFGAKVKSSAP 320
Tetraodon       WFQNRRVKEKKFVAKVKSSAP 320
Fugu            WFQNRRVKEKKFVAKVKSNA- 319
Danio           WFQNRRVKEKKFIAKVKNTAP 303
                ************ ****..*
```

Fig. 1. Clustal W alignment of *hoxb13a* vertebrate paralogs. Spheroides = *Spheroides nephelus hoxb13a* reference protein (AAQ72843). *Spheroides* is the Southern pufferfish (Amores et al. 2004). Tetraodon = *Tetraodon nigroviridis hoxb13a* (Ensembl: Un_random 38176495-38178237) and Fugu = *Takifugu rubripes hoxb13a* (Ensembl ID: SINFRUG00000124866). Danio = translation of predicted genomic coding sequence for *Danio rerio hoxb13a*. Key: (*) = identical in all sequences; (:) = conserved substitutions; (.) = semi-conserved substitutions.

formed using Superscript III (Invitrogen). The resulting cDNA was used as a template for the PCR reactions at about 100 ng/25 µl reaction. The PCR mix consisted of 1 U of *Taq* DNA polymerase (Qiagen, Venlo, The Nethrlands), 0.8 pmol of each primer, and 0.2 mM dNTPs. PCR products were cleaned using Qiagen columns.

Sequence reactions were carried out with the BigDye Terminator Cycle Sequence Kit (Applied Biosystems, Foster City, CA, USA) in a 10 µl reaction volume with a 2 µl reaction mix and a variable primer concentration depending on the concentration and size of the input PCR sample. Sequence products were cleaned with Sephadex columns (Amersham Biosciences, Little Chalfont, UK) and run on an ABI 377, and edited with Sequencher (Genecodes, Madison, WI, USA).

Cloning and sequencing of *hoxb13*

Hoxb13a PCR products were inserted into the pCR II-TOPO and pCR 4-TOPO vectors provided in the TOPO TA cloning kits (Invitrogen). Sequencing was performed by ServiceXS (Leiden, The Netherlands) on both strands of the vector using M13 forward and reverse primers.

In situ hybridization

Zebrafish embryos (*Tübingen* strain) were processed for in situ hybridization following the high resolution whole-mount in situ hybridization as described previously (Thisse et al. 1993). For hybridization with *hoxb13* probe, eight embryos of 24 hpf ($n = 8$) were used. Samples were treated with proteinase K (10 µg/ml) for 10 min. Color reactions were developed with the NBT/BCIP substrate (Roche, Almere, The Netherlands).

Microarrays

Protocols were as described previously (Meijer et al. 2005). Briefly, dual-color hybridizations were performed with RNA probes from control fish versus *Mycobacterium*-infected fish. Both comparisons were performed in duplicate, resulting in a total of four data sets for the MWG array and four data sets for the Sigma array. To check for consistency within the spotted oligonucleotide microarrays, Ambion spikes and oligo test sets were compared. For the Affymetrix analysis, two GeneChips were used for the control fish and two GeneChips for the infected fish.

Data analysis

For spotted oligonucleotide arrays, individual feature intensities were extracted from scanned microarray images using GenePix Pro 5.1 software (Axon Technologies, Westburg B.V., Arnhemseweg, The Netherlands). Affymetrix GeneChip data were extracted and normalized using Affymetrix GCOS software. Data outputs were imported into Rosetta Resolver 4.0 (Rosetta Inpharmatics LLC,

Kirkland, WA, USA). Individual arrays were normalized using default settings. For spotted oligonucleotide data, the intensity data from each channel were processed with the Ratio-splitter without a common reference tool. Data thus obtained were centered, scaled, and combined statistics were calculated using the Default Intensity Experiment Builder. Affymetrix data were centered, scaled, and combined statistics were calculated using the Affymetrix Intensity Experiment Builder (for details, see http://info.rosettabio.com/).

RESULTS

By blasting the zebrafish genome with hox reference proteins from zebrafish and other vertebrates, we recovered complete or partial candidate sequences for all previously described zebrafish *hox* genes, as well as a new zebrafish gene, *hoxb13a* (see Table 2). Sequence mismatches between the genomic and reference sequences are shown in Table 3.

Genomics

The arrangement of zebrafish *hox* genes on genomic contigs is shown in Fig. 2. Homology assumptions are confirmed by the arrangement of each on the appropriate cluster with the exception of *hoxa11a*, *a13b*, *b3a*, *b10a*, and *c13b*, which were found in different genomic contigs, or were otherwise in anomalous locations that we assume are because of errors in assembly. The alternative transcript of *hoxb3a*, and the unusually positioned exon 1 of this gene (Hadrys et al. 2004), were not confirmed here. This is likely because of extensive areas of bad sequence between *hoxb5a* and *b4a*, and assembly errors indicated by the swapping of positions of exons 1 and 2 of *hoxb3a* in Zv.4.

The *hoxaa*, *bb*, and *d* clusters were each recovered complete on a separate contig. The *hoxaa* cluster included the *hoxa10a* pseudogene (Amores et al. 1998; GenBank NG_001593). *Hoxa11a* was on the negative strand in Ensembl. *Hoxa13b* was on a contig different from the other *ab* genes. *Hoxa9b* exon 2 was on the negative strand, and exon 1 was misplaced to the 3′ end of the cluster.

All sequences of the *hoxba* cluster recovered were located on the same contig with the exception of *hoxb10a*, the only trace of which was a fragment of exon 2 on a separate contig. For the *hoxca* cluster, all genes except *hoxc13a* exon 1 were recovered on the same contig. The complete ORF for *hoxc13a* was located on a separate contig. All *hoxcb* members were on the same contig, except *hoxc13b*.

RT-PCR and sequencing

Primers were designed using our ORF predictions (see Table 2). Reverse primers corresponded to sequences near the beginning of exon 2. This was in order to avoid, as much as possible, the highly conserved homeobox. Forward primers were chosen from between 200 and 300 bp upstream of the predicted end of exon 1. As can be seen in Table 1, products were obtained for all genes analyzed except *hoxa13b*, *b1a*, *b4a*, and *b9a*—and two for which the genomic sequence was incomplete, preventing us from designing primers (*hoxb3a* and *b10a*). However, published ESTs were available for all of these genes, and so we were able to verify annotations, and expression was verified in the microarrays.

Phylogenetic analysis

In order to validate our assumptions about the homology of the *D. rerio* annotations, we compared translations of the *Danio* genomic sequences with corresponding translations of genomic sequences from humans (Fig. 3). The sequences were aligned using AlignX (Vector NTI 9.0, Invitrogen), and analyzed using the UPGMA algorithm implemented in MEGA version 3.0 (Kumar et al. 2004). Further analyses of exons 1 and 2 amino acid sequences, and the complete CDS, using both likelihood (TreePuzzel 5.2) and parsimony (PAUP v4), produced similar results (the trees are posted on our website, www.mk-richardson.com).

Sequences that showed significant mismatches with reference proteins were excluded from the analysis, in case errors led to a false placement of the sequence. The resultant tree was well resolved, recovering all the expected paralog groups with the exception of the paralog 6 group, which formed a polytomy with the 4, 5, and 7 paralog groups. Notably, *Danio hoxa11a*, *a13b*, and *b13a* were placed within the appropriate paralog group, despite their incongruous positions on the genome.

A new gene: *D. rerio hoxb13a*

Although there was no *Zfin* accession for this gene and no zebrafish reference sequences, the translated genomic sequence showed strong homology with reference proteins from paralog group 13 in the zebrafish and other species (Fig. 1). The 3′ end of exon 1 differed from the multi-species *hoxb13* consensus. However, the gene is expressed in vivo as shown by in situ hybridization (Fig. 4). Expression can be seen in the distal part of the tail at 24 hpf embryos. The gene had an entry on Ensembl as an unknown transcript (Ensembl ID: ENSDARG00000010288).

The predicted protein was blasted against genome builds for *Takifugu rubripes*, (Ensembl release 27.2d.1) and *Tetraodon nigroviridis*, (Ensembl release 27.1b.1). Putative *hoxb13a* homologs were recovered from *Tetraodon* (Un_random 38176495 38178237) and *Takifugu* (Ensembl ID: SINFRUG00000124866) genomes.

Microarrays

We have analyzed mRNA isolated from adult zebrafish using three different platforms (Affymetrix Genechips, as well as

Table 2. Summary of genes recovered as genomic sequences, their Zfin names (see http://zfin.org/), Ensembl (Zv4) I.D. (= ENSDARG000000+number), the GenBank accession numbers of our PCR products, translations of the 5′ and 3′ ends of our predicted ORF, and the amino acid sequences at our predicted splice sites

Zfin gene	Ensembl I.D.	GenBank accession	Start	Exon1–intron	Intron–exon2	End
Aa						
hoxa1a	04438	DQ060531	MSTFLDFS	VKRNPPKT	GKAGEYGF	TVEAYSSN
hoxa3a	10987	DQ060532	MQKATYCD	QKSCSIIS	VESCAGDK	EAPKLTHL
hoxa4a	20687	DQ060533	MIMSSYLI	MKKVHVNT	VTASYSGG	PTPCSSNL
hoxa5a	01784	DQ060534	MSSYFVNS	MRKLHISH	DNLAGPEG	AAGSGYRP
hoxa9a	09461	DQ060535	MSTSGALT	EGKPGADP	ENPVSNWL	NKNETKED
hoxa11a	09045		MMDFDERV	YMLFYKRI	GGPRFRKK	YYSTNPLL
hoxa13a	07609		MTTSLLLR	NVWKSSIP[1]	ESVSHGGA	VNKLKSSS
Ab						
hoxa2b	23031	DQ060536	MNYEFERE	GPLYFSPQ	GSPEISDG	IDLQHLSY
hoxa9b	23013[2]	DQ060537	MSTLGTLS	ETKLDLDP	NNPSSNWL	KMKKCNKD—[3]
hoxa10b	31337	DQ060538	MSCSDSPS	EKAVTVTK	AGDSKSES	LSANFSFS
hoxa11b	07009	DQ060539	MMDFDERV	EDKFSGSS	NGQKTRKK	YYTTNPLL
hoxa13b	02503		MTASLLLH	LWKSSIQG	TDGASVRR	VNKYKGIS
Ba						
hoxb1a	08174		MDSSRMNS	RNPPKTG	KVAEYGLG	EASPSPDS
hoxb2a	00175		MNYEFERE	(genomic sequence mismatches)		IDLQHLQF
hoxb3a	29263		MQKTTYYD	SPSASSAN	AESSGGEK	EAPKLTHL
hoxb4a	13533		MAMSSYLI	MKKVHVNI	VSPNYSGG	ASGPPPSL
hoxb5a	13057	DQ060540	MSSYFVNS	MRKLHISH	DMTGPDGK	TAGSAFQP
hoxb6a	10630	DQ060541	MSSYF(L/V)NS	MQRMNSCN	GTFGNAGR	EEEEKRTE
hoxb7a	00193	DQ060542	MSSLYYAN	YPWMRSTG[4]	ADRKRGRQ	DEEEEDDE
hoxb8a	14115	DQ060543	MSSYFVNS		PVAAGRRR	DCDKAKQM
hoxb9a	01753	DQ060544	MSISGTLS	EDKEGPDQ	DDPSANWL	MNKDQPKE
hoxb10a	11579[5]		(bad/missing sequence)			DPGTSFTV
hoxb13a[6]	10288	DQ060545	MTTSLVLN	HLWKSALA	DVVAHQHD	AKVKNTAP
Bb						
hoxb1b	01353	DQ060546	MNSYLDYT	VKRNPPKT	VKVAEYGI	TSDSSTAI
hoxb5b	05395	DQ060547	MSSYFLNS	MRKLHISH	DMTGPDGK	TAGSAFQN
hoxb6b	26513	DQ060548	MSSYFVNS	MQRMNSCN	GMPGSTGR	EEDGGKAG
hoxb8b	01254	DQ060549	MSSYFVNS	LFPWMRPQ	ATGRRRGR	SEASSNSK
Ca						
hoxc1a	20383	DQ060550	MNSYHGFR	VRRNQSRA	AKIQLGKC	SDTCSSPD
hoxc3a	05235	DQ060551	MNNNSFHE	SSINAMES	GDSKYSNG	YETPSMNW
hoxc4a	06210	DQ060552	MIMSSYLM	MKKIHVST	VNSSYNGA	RGEDITRL
hoxc5a	19216	DQ060553	MSSYVGKS	MTKLHMSH	ESDGKRSR	KLKVKGGL
hoxc6a	17517	DQ060554	MNSYFANP	MQRMNSHS	GVGYGSDR	EEEPKKKD
hoxc8a	16658	DQ060555	MSSYFVNP	MFPWMRPH	APGRRNGR	EEKEESKE
hoxc9a	13621	DQ060556	MSATGPIS	DDKAELDP	NNPVANWI	EKNDSKEQ
hoxc10a	20576	DQ060557	MSCPNNVA	DDSESELK	DESKLEKA	ELTGYSFN
hoxc11a	28655	DQ060558	MFNSVNLG	NASKSSHS	IPRAQERR	YFSGNPLL
hoxc12a	18127	DQ060559	MGEHNLLN	ASNIAGGG	GAPWYPMH	REQALSFF
hoxc13a	exon 1: 19821 exon 2: 28639	DQ060560	MTTSLVLH	HLWKSPFP	DVVPLQPE	KTNNHMHT
Cb						
hoxc6b	13531	DQ060561	MNSYFTNP	QRMNSHSG	VGYGSNKR	TAEKDEHD
hoxc11b	27686	DQ060562	MFNSVNIG	NQTKSGHS	TTPRMRKK	YFSGNPLM
hoxc12b	04682	DQ060563	MGEHNLFN	TSVAALNG	GALWYPMH	REQALSNF
hoxc13b	13448		MEGLSGNC	HLWKSQFS	DVVPHQAE	SSTNMHSV
Da						
hoxd3a	33148	DQ060564	MQKATYYD	KSTNCPAA	GETCDDKS	DAPKLTHL
hoxd4a	10540	DQ060565	MEGGKKDN	MKKVHVTT	VNPDYTGP	SQTEITTL
hoxd9a	27006	DQ060566	MSTSSALS	KQQQQLDP	SNPAANWI	RERSSKDP
hoxd10a	16874		MSFPNSSP	ELPHREGK	AESKNDTP	LTSNLTFS
hoxd11a	17558		MTDYDDRN	DEEKNSGS	SATKSRKK	YFTGNPLF
hoxd12a	18023		MCEHNLLS	DPSAIDT[7]	GLPWCPSQ	REHTFTIY
hoxd13a	26670		MDGGGLDE	HIWKPSLT	EEAAAASF	RPDVCIKC

Notes:
[1]The boundary given in the table is taken from AY303229 because the genomic sequence contains mismatches after SSYASSPY.
[2]The Ensembl I.D. given is for exon 2 only; there was no Ensembl I.D. for exon 1 but its GENSCAN I.D. was 00000040218.
[3]Protein alignments suggest that there should be four further terminal amino acids after KMKKCNKD.
[4]Exon 1 not recovered.
[5]Exon 1 not recovered; no full-length *Danio* EST found.
[6]*Hoxb13a* is not listed in Zfin.
[7]The genomic region 5′ to DPSAIDT contains sequence mismatches.
ORF, open reading frame.

Table 3. Mismatches in reference proteins, reference DNA, and Zv4 genomic sequences for those genes that could recover our own ESTs (see Table 2) or for which sufficient number of independent sources of sequence were available

Zfin gene	Proteins	DNA	Zv4 genomic sequences
hoxa3a	(NP_571609.1); p.D154delinsRQ; p.AS273_274HL	(NM_131534.1); g.459_460insT; g.473_474insAC; g.818delG; g.823delGinsTA	
hoxa4a	(AAD15939.1) p.RSSSSAPSNHHVETDATQQ214_232APRALHPPTITWKRTPLS	(AF071246.1); g.641delC; g.696_697delAG	g.677G>A (Y13947.1 and AF071246.1)
hoxa5a	(NP_571615.1) p.1_38del MSSYFVNSFCGRYPNGVDYPLHN YGDHNS SGQCRDSTG: Actual protein start is 38 aa. upstream of the reference protein start		ENSDARG00000001784 5′ truncated; annotation from NP_571615.1
hoxa9a	(NP_571607.1) p.KPGAD166_170NRALI	(NM_131532.1) g.495delG; g.510_511insT	
hoxa2b			g.892G>A (NM_131106.1 and Y13945.1)
hoxa9b	(NP_571608.1) p.I15L; p.175_176insSKCNQ: wrong splicing site	(NM_131533.1) g.43A>C g.390G>C g.392T>G g.524_525ins GTAAGTGCAACTAAG: wrong splicing site	exon 1 misplaced to 3′ end of cluster; exon 2 on the negative strand, lacks stop codon; g.136A>G (NM_131533.1 and CAD59109)
hoxa10b			g.275G>A (Q8AWY2 and CAD59110)
hoxa11b	(NP_571222.1) p.M269L		
hoxb1a	(NP_571190.1) p. MDSSR1_5delinsMGYEQFR; p.SFL8_10LSW; p.Q70R	(NM_131115.1) g.1insATGGGGT; g.4G>A; g.13A>G; g.16delA; g.21delC; g.32insG; g.45T>C; g.209A>G; g.784C>T	
hoxb2a			g.9_10insT; g.15_16insT; g.19_20insA; g.31T>G; g.384C>T; g.402delC; g.405_406insA; g.440_621del : missing beginning exon 2; bad sequence region. (NM_131116.2)
hoxb3a			exon 1 downstream of exon 2; g.494_640del : missing beginning exon 2; bad sequence region. (NM_131117.2)
hoxb6a	(NP_571194.1) p.V6L	(NM_131119.1) g.16G>C	
hoxb7a			stretch of bad sequence in intron
hoxb8a			exon 1 missing.
hoxb9a	(NP_571196.1) p.GPDQ168_171DRIKVSYNLG	(NM_131121.1) g.503delG; g.514_515insGTAAGTTATAA-CCTAGGAA : longer exon 1	g.474C>G
hoxb10a			exon 1 missing; exon 2 isolated in a different contig than the rest of the cluster

Table 3. (Contd.)

Zfin gene	Proteins	DNA	Zv4 genomic sequences
hoxb1b	(NP_571217.1) p.Y158C	(NM_131142.1) g.473A>G	
hoxc3a			g.277A>G
hoxc5a	(NP_571219.1) p.RTDDIKMETTSAI122_134del insSRRYQNGDYFSD	(NM_131144.1) g.364C>A; g.368_369delCA; g.404delA	g.420T>G
hoxc10a	(AAD15950.1) Reference protein incomplete p.AT254_255QR; p.R268X; p.Y271N	(AF071257.1) g.740_747delATGAGTCT insCGGGCAAGAGTGA AGTTCTAA; g.754_755insGA; g.768_769insG; g.802A>N; g.811T>A	
hoxc13a	(NP_571618.1); p.F154L; p.A175V	(NM_131543.1) g.460T>C; g.524C>T	Other mismatches found, but no third sequence source is available.
hoxc6b	(NP_571605.1) p.M109V; p.S114V; p.SG128_129delinsR; p.Q138R; p.Q145R	(NM_131530.1) g.325A>G; g.340_341AG>GT; g.369G>C; g.383_385delGTC; g.413A>G; g.434A>G;	g.298delG; g.403G>A; g.407G>A
hoxc12b	(NP_571620.1) p.K118G	(NM_131545.1) g.352_353AA>GG; g.471A>T;	g.333delC; g.667_668insN; g.669delA; g.721A>C; g.723_724insC; g.741A>T; g.743_744insT; g.748G>T; g.754_755CT>TA; g.769C>A; g.771A>C; g.788T>C; g.799C>G; g.827G>A;
hoxd3a	(CAA74286.1) Frame shift -1 from p.287 onwards	(Y13948.1) g.861delG	
hoxd4a	(NP_570113.1) p.G125C; p.I199S; p.C204S	(NM_130757.1) g.373G>T; g.417G>C; g.598T>G; g.612G>C	
hoxd9a	(NP_571201.2) p.P112S; p.D154Y	(NM_131126.2) g.334C>T; g.460G>T	
hoxd11a	(CAA61030.1) p.1_13delMTDYDDRNNCASN; p.G198X	(X87751.1) Shorter ORF: g.1_39delATGACGGACTA CGATGATCGCAACAA CTGTGCATCTAAT; g.372C>T; g.592G>N	DQ069272
hoxd12a	p.IAPFQPSLSAQNIRPAFTD 165_183delins SXSIPAVTERPEYQTS FHRWYAQLLQDPSAIDT (frame shift and mutations at the end of exon 1)	No DNA seq. available	DQ069273

Hoxa11a, hoxa13a, hoxa13b, and hoxd10a are therefore not listed, although there are mismatches between Zv4 genomic sequence and protein or DNA reference sequences because the polarity of changes could not be clearly assessed. Position+1 of CDS (g.) or first amino acid of the translated protein (p.) used as a numbering reference. Table entries given for wrong sequences according to our analysis. Polarity of changes shown always begins with our corrected sequence. In protein and ESTs columns, accession numbers indicate sources that contain mismatches as indicated. In Zv4 genomic sequence column, accession numbers reflect sources for genomic annotation; genomic sequences contain mismatches as shown.

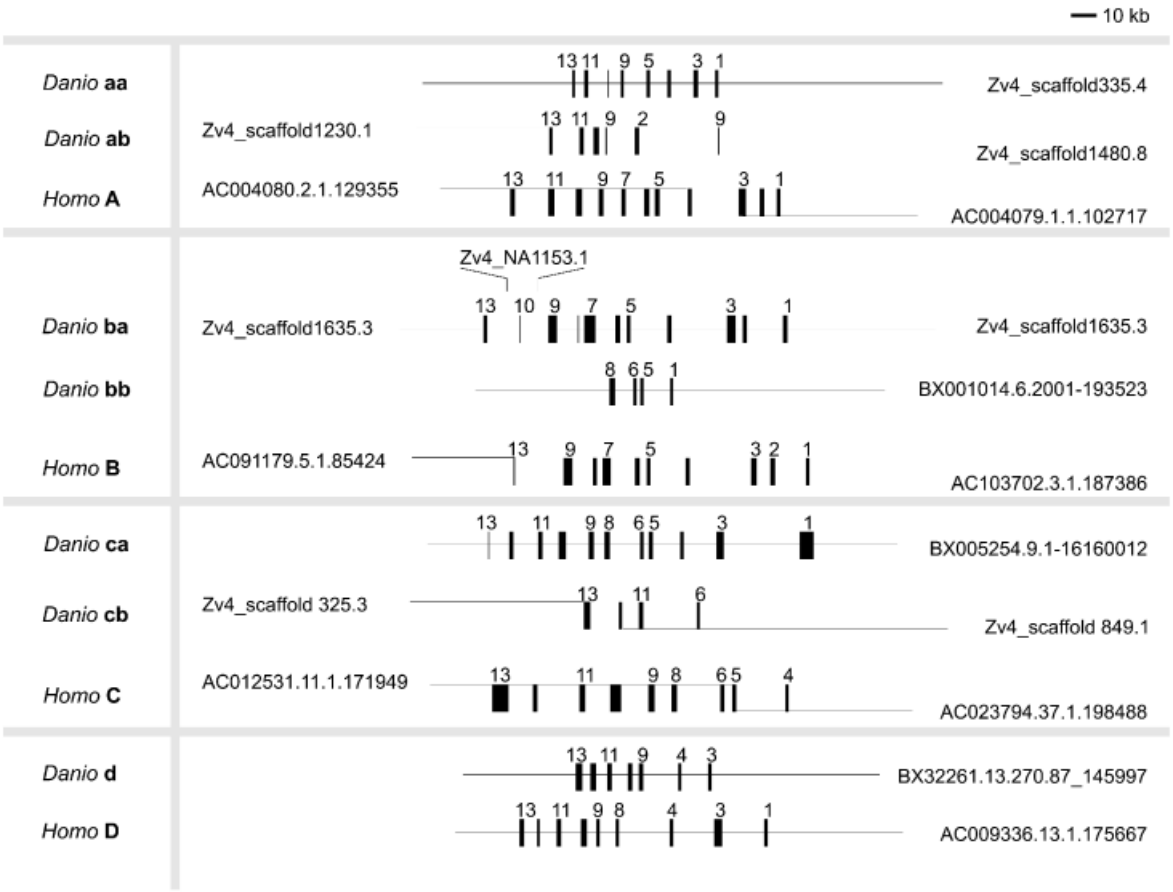

Fig. 2. Genomic map of the zebrafish *hox* clusters (Ensembl Zv4) compared with those for humans (*Homo sapiens*; NCBI build 34 version 3). Code numbers are contigs. The human annotations were checked against reference proteins and ESTs, and *Danio* annotations were corrected using the evidence listed in Table 1. *Notes: Danio hoxa11a* was on the complementary strand in Ensembl. *Danio hoxa10a* is a pseudogene. *Danio hoxa9b*: two exons are shown because exon 2 was on the negative strand but in the expected location, whereas exon 1 was misplaced to the 3′ end of the cluster. *Danio hoxb8a:* only exon 2 is shown; exon 1 was not found. There is a long stretch of bad sequence between *Danio hoxb5a* and *b4a*. *Danio Hoxb3a* exon 2 was incomplete and misplaced. *Danio hoxb10a*: only exon 2 was found and was on a contig separate from the rest of the cluster members. Only exon 2 of *Danio hoxc13a* is shown in the figure; exon 1 was on a different contig (AL935205.13.1-150641).

spotted oligonucleotides from Sigma-Genosys, Zwijndrecht, The Netherlands and MWG-Biotech AG, Ebersberg, Germany). Having first performed a systematic annotation of the entire complex, we were able to identify correctly all the *hox* genes in the microarray data sets as supplied by the manufacturers. Various *hox* genes were mis-annotated; we provide a data set with corrected *hox* identities on our website, www.mk-richardson.com.

Our analysis allows us to conclude that all but two of the total 49 *hox* genes are present in those datasets. The two missing genes are *hoxc12a* and *hoxb13a* because no public ESTs were available; *hoxb13a* is described here for the first time. A previous study (Meijer et al. 2005) analyzed the effect of *Mycobacterium* infection in adult zebrafish. Using their data, we were unable to find any differential expression of the *hox* genes. Therefore, all further analyses were performed with data from both infected and noninfected fish.

Average intensities and their respective errors were obtained for the 47 *hox* genes present in any of the three platforms used. We could detect significant levels of expression in adult tissues of all *hox* genes except for *hoxa5a*, *hoxc1a*, and *hoxc6a* (Table 4). For these three genes, as well as for the two not represented in the data set, gene expression in adult zebrafish was detected by RT-PCR (data not shown).

Comparison of the results for different oligonucleotide sequences designed for the same gene shows that the observed expression level depends greatly on the probe design. This is especially relevant for Affymetrix Genechips where data from 16 individual 25-mers are combined to obtain the final intensity output. *Hox c11a* was found to be expressed at high levels with MWG and Sigma probes. However, with Affymetrix, the level of expression detected was substantially lower. Analysis of the individual oligonucleotide data for *hoxc11a* showed that although the average signal is low, several oligos gave

Fig. 3. UPGMA cladogram of amino acid translations of full-length genomic *hox* sequences of *Danio rerio* (Dr) and *Homo sapiens* (Hs). Our homology assumptions about the *Danio rerio* genomic sequences are broadly supported. Isolated anomalies were noted under other search methods; see additional trees on our website at www.mk-richardson.com

relatively high signals (Fig. 5). This exemplifies the existence of statistically significant differences for different probes of the same gene. These results highlight the importance of having genomic sequences, taking into account single-nucleotide polymorphisms, for good oligonucleotide design.

In the Affymetrix set, several genes had probes designed in the anti-sense strand. In our data, we found that there was significant expression of the anti-sense oligonucleotides for all the *hox* 8 paralogs (i.e., *hoxb8a*, *hoxb8b*, *hoxc8a*) as well as for *hoxb7a*. Interestingly, this confirms that anti-sense ESTs for these genes indeed represented anti-sense expression in vivo rather than sequence orientation errors.

DISCUSSION

Quality of genomic data and reference proteins

The zebrafish genome project is providing a valuable new resource for biologists. Not only has immense progress been made in a relatively short span of time, but the sequence quality, and accuracy of annotations, is improved with each new release. The depth of the most recent genomic build (Zv4) is high, in the sense that we were able to recover most *hox* genes, as well as a novel gene, *hoxb13a*. A few anomalies were noted.

Hox genes are well known to appear in closely packed clusters with relatively small intergenic distances. This information could be very useful to determine the quality of the genome sequence available. We have found a number of regions with either poor-quality or missing sequence, or gaps (Table 3). Some regions were present in duplicate, but with sequence differences between the duplicates.

Examples of such anomalies include the inversion of *hoxa9b* exon 2 and the displacement of its first exon to an anomalous location; the absence of *hox*b10a from its cluster with its second exon only being located on a different contig; and the presence of *hoxc13a* exon 2 on two different contigs. Such anomalies are possibly because of errors in genome builds for shotgun sequences. The combination of genomic sequences and PCR also allowed us to note mismatches with reference proteins (Table 3).

Homology assumptions

Overall, the assumptions about the homology of our Zv4 genomic annotations appear to be valid when a comparison was made with human sequences by phylogenetic analysis. Particularly under parsimony, all the trees are roughly split between anterior and posterior groups (Fig. 3 and additional cladograms on our website, www.mk-richardson.com).

This is further confirmed by the fact that our predicted *Danio* genes are arranged in the appropriate colinear order within the clusters (Fig. 2). When CDSs are compared within each paralogous group, it is seen that the relative sizes of exon 1 versus exon 2 are very similar between *D. rerio* and *Homo*

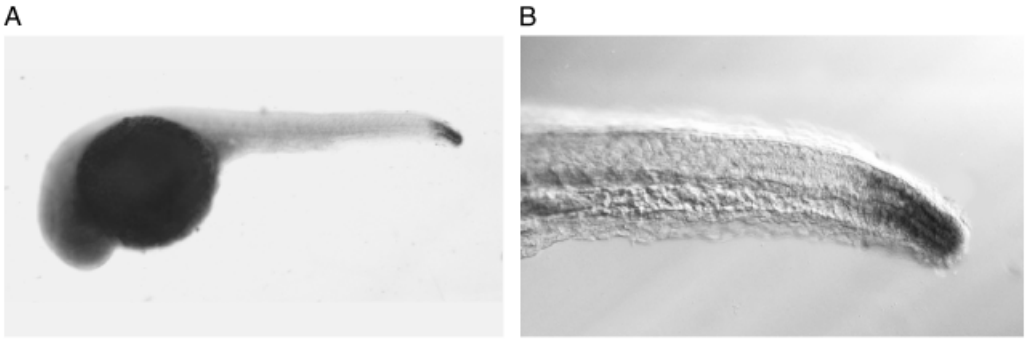

Fig. 4. Whole-mount in situ hybridization with *Danio rerio hoxb13a* probe in 24 hpf zebrafish embryo. Note the hybridization at the tip of the tail. Caudal is to the right and dorsal is to the top. (A) Whole embryo, left lateral view. (B) Detail of tail.

sapiens (Fig. 6). Notable exceptions include the enlargement of exon 1 in *Homo sapiens* by polyalanine tracts and other amino acid repeats (e.g., in *HOX*D13, A13, and A11; Utsch et al. 2002; Lavoie et al. 2003).

Hoxb13a

Hoxb13 has been described in tetrapods (Zeltser et al. 1996; Carlson et al. 2001), and a homolog is present in teleosts (*T. rubripes* and *Spheroides nephulus*; see Amores et al. 2004). It was thought to have been lost in the lineage leading to zebrafish because previous characterizations of the *hoxb* clusters failed to find it. Both pufferfish species mentioned have two *Hoxb* clusters, but only one of these in each species contains a 13 paralog. Although many studies have been carried out to determine *hox* cluster organization in the zebrafish, the absence until recently of a full genome sequence made it impossible to confirm whether the full set of *hox* genes had been discovered.

We have examined the expression of *hoxb13a* during zebrafish axis development. As in mouse (Zeltser et al. 1996) and axolotl (Carlson et al. 2001), the expression is restricted to the tip of the developing tail at 24 hpf stages (Fig. 1B). This finding is consistent with the paradigm of expression colinearity and resembles the pattern of other *D. rerio* 13 paralogs, such as *hoxc13a* and *hoxc13b* (Thummel et al. 2004). We are currently examining the possible role of this gene in zebrafish fin regeneration, because posterior *hox* are implicated in this process (Geraudie and Borday 2003).

Microarray analysis

We have shown that the 49 *hox* genes are expressed in adult tissues. Our microarray analysis of adult fish was able to detect that 44 out of the 47 present in the oligonucleotide sets are expressed. This is the first report where expression has been demonstrated for nearly all members of the zebrafish *hox* family excluding *b13a*. We conclude that with improved oligo design and calibration technology, a robust quantitative analysis of expression will be possible.

Our validation of the oligonucleotide annotations in all three platforms will be of great value for future microarray analyses of *hox* gene expression. Further, they point to a need for validation of sequence in microarray studies in general. Considering the varying results that we obtained for the same gene with different oligonucleotides (e.g., see Fig. 5 for Affymetrix data for *hoxc11a*, and Table 4 for an overall comparison between technologies), we are developing a new set of oligonucleotides based on our genomic annotations and PCR. These will be used for expression profiling of *hox* genes in different tissues at different stages of development.

All available platforms are based on EST database information from a variety of zebrafish strains. However, the high number of SNPs in zebrafish, together with the mismatches described here between the ENSEMBL genomic sequence and previously published reference sequences, highlights the importance of using the most accurate sequence sources for oligonucleotide design. This is particularly important for technologies such as Affymetrix, where shorter oligos are used and mismatches may strongly affect the hybridization. A further issue is the importance of recording the zebrafish strain used to prepare specific genomic resources.

Here, we also present evidence of expression of the antisense of the hox8 paralog group and *hoxb7a*. Although antisense *hox* sequences are present in the EST database, they come from incomplete cDNA clone sequencing projects and annotation is insufficient to conclude much about their pattern of expression. It would be of great interest to determine whether these anti-sense mRNAs play a role in the regulation of *hox* genes during zebrafish development. Together with previous reports of in vivo anti-sense expression of *hoxA11* (Hsieh-Li et al. 1995) and *hoxD3* (Bedford et al. 1995) in other species, the expression of the anti-sense transcripts reported here suggests that they could provide a mechanism for the regulation of *hox* gene expression during development.

The presence of anti-sense transcripts in the mRNA pool also has implications for the detection techniques based on nucleic acid hybridization such as in situ hybridization, microarrays, and qPCR, because of competition of the anti-sense strand with the probes used for the detection. Possible modulation of detection kinetics should be taken into account when performing quantitative studies, if possible. Therefore, the existence of a larger set of *hox* genes regulated by in vivo

Table 4. Results of microarray analysis; mRNA from adult zebrafish was used for the experiments

	MWG		Sigma		Affymetrix	
Zfin gene	Intensity	Error	Intensity	Error	Intensity	Error
hoxa1a	**125**	**36**	30	15	**15**	**3**
hoxa3a	**308**	**61**	12	9	**66**	**6**
hoxa4a	**76**	**15**	36	26	**11**	**3**
hoxa5a	10	50	26	16	− 1	2
hoxa9a	73	16	10	13	**14**	**2**
hoxa11a	56	7	**58**	**17**	**16**	**3**
hoxa13a	63	7	**129**	**45**		
hoxa2b	**117**	**46**	**48**	**19**	**12**	**2**
hoxa9b	**104**	**22**	**180**	**68**	**13**	**3**
hoxa10b	**750**	**266**	**74**	**21**	**13**	**3**
hoxa11b	**84**	**24**	36	14	**12**	**3**
hoxa13b	66	10	25	11	**13**	**2**
hoxb1a	**81**	**15**	2	12	**9**	**1**
hoxb2a	**271**	**177**	27	18	**31**	**5**
hoxb3a	**140**	**32**	29	20	**78**	**4**
hoxb4a	57	8	14	10	**45**	**5**
hoxb5a	**234**	**164**			**51**	**7**
hoxb6a	65	15	72	23	**17**	**2**
hoxb7a	**577**	**83**	**78**	**23**	**65**	**5**
hoxb7a AS					**22**	**4**
hoxb8a	**206**	**24**	23	16	**89**	**7**
hoxb8a AS					**38**	**3**
hoxb9a	**135**	**18**	**41**	**14**	**34**	**7**
hoxb10a	**110**	**13**			**16**	**3**
hoxb1b	**77**	**18**	40	41	6	2
hoxb5b	**90**	**13**	10	11	**17**	**2**
hoxb6b	**169**	**37**	**725**	**295**	7	2
hoxb8b	**78**	**17**	29	34	**11**	**2**
hoxb8b AS					**25**	**4**
hoxc1a	36	7	82	104	− 14	2
hoxc3a	**83**	**22**	**66**	**19**	**26**	**3**
hoxc4a	**81**	**19**	30	16	**9**	**1**
hoxc5a	**237**	**52**	34	22	2	1
hoxc6a	69	10	− 1	11	2	2
hoxc8a	**104**	**20**	**44**	**10**	**26**	**2**
hoxc8a AS					**20**	**3**
hoxc9a	**75**	**11**	**62**	**24**	**62**	**5**
hoxc10a	**2393**	**500**	60	31	**43**	**5**
hoxc11a	**1620**	**570**	**3052**	**418**	**24**	**3**
hoxc13a	**73**	**9**	**211**	**42**	**38**	**4**
hoxc6b	**69**	**7**	21	14	**11**	**2**
hoxc11b	**1862**	**806**	− 3	11	**11**	**3**
hoxc12b	**286**	**66**	3136	**817**	**18**	**3**
hoxc13b	**101**	**16**	38	15		
hoxd3a	**252**	**79**	55	41	3	1
hoxd4a	**79**	**16**	28	15	− 8	3
hoxd9a	**95**	**13**	8	21	**12**	**4**
hoxd10a	**118**	**33**	**129**	**27**	**33**	**2**
hoxd11a	**95**	**8**	9	12	**42**	**4**
hoxd12a	**76**	**17**	**35**	**16**	5	2
hoxd13a	**85**	**18**	− 4	10	7	1

MWG, Sigma, and Affymetrix refer, respectively, to the three microarray technologies used (see Materials and Methods for details). Average intensity and error (1 SD) are given. Bold indicates that expression level was found to be significant ($P \leq 0.01$). Expression of the underlined genes was not significant for any of the three platforms. As can be seen, all but three of the 47 genes listed show significant levels of expression in adult tissues. Gene names are from http://zfin.org/.

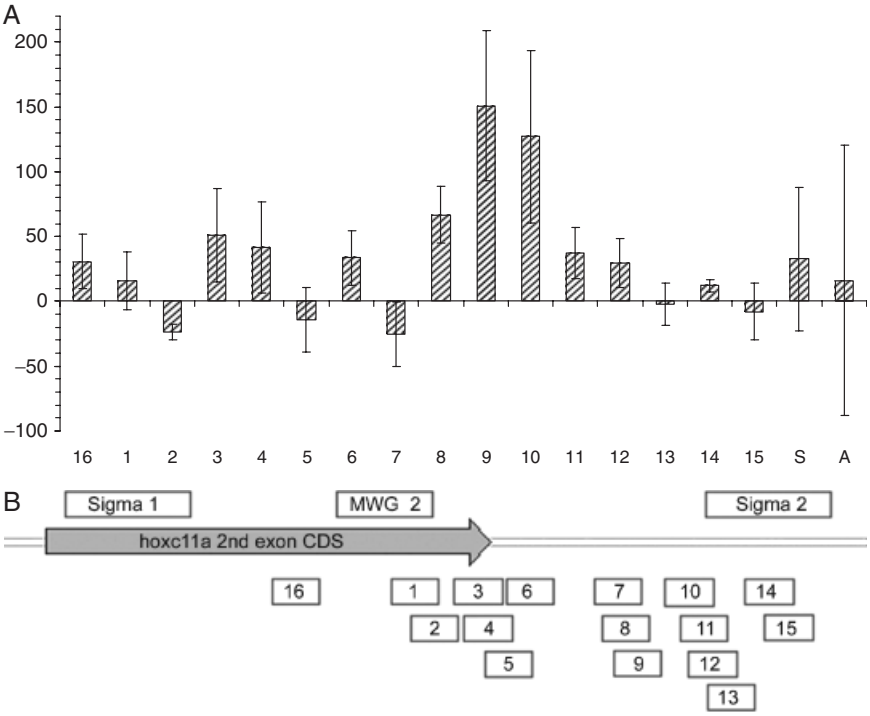

Fig. 5. (A) Difference of average intensities (± 1 SD) for each perfect match/mismatch pair of probes recorded for *hoxc11a*, ordered according to their genomic localization, 5′ to the left. S represents the average of the differences for all sense probes; A represents the average of the differences for all anti-sense probes; (B) genomic localization of the 16 Affymetrix, 2 Sigma and 1 MWG oligonucleotide probes for *hoxc11a*. For some of the oligos showing no expression, we found SNPs (data not shown). It is therefore helpful to have genomic sequence data for oligo design.

anti-sense transcription could not only be a source of further knowledge regarding the mechanisms of *hox* genes regulation but may also have implications for the data recovered so far. These issues could be usefully examined further.

As comparative and functional genomics become increasingly important in evolutionary developmental biology, there will be a growing need for high-quality annotated genomic sequences. It is widely recognized that manual annotation is extremely labor intensive and time consuming. However, we have shown here that it is important not to rely entirely on automated annotations. Here, we have illustrated this point with analysis of the *Hox* gene family, and have shown that sequence errors and misannotations do exist in public and commercial resources. The validity of the conclusions of any genomics study critically depends on the quality of the annotation of the underlying data.

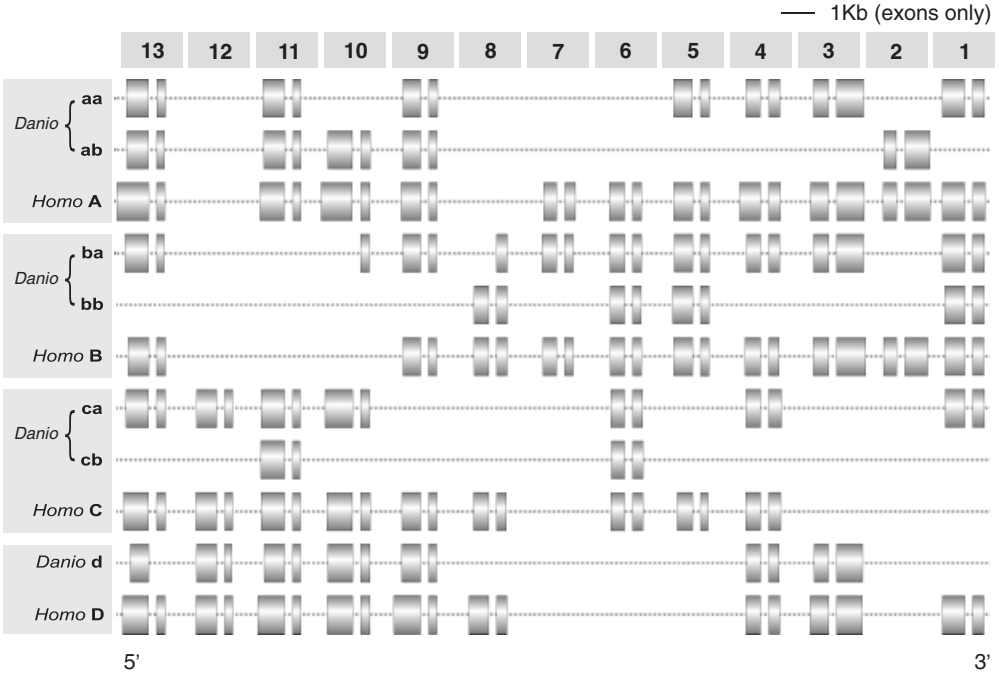

Fig. 6. Diagram showing proportional size comparison of *hox* gene exons in zebrafish (*Danio rerio*) and humans (*Homo sapiens*). Paralogous groups are given in columns with the number indicated at the top. The introns are all shown at the same fixed size to allow exon size comparison to be made. Intergenic distances are not to scale. Note the larger size of exon 1 in some human genes (e.g., HOXA13).

Acknowledgments

We thank Enrique Salas-Vidal and Eric Wielhouwer for helpful discussions. M. C-A. was supported by the University of Leiden. H. P. S. and A. H. M. were supported by a European Commission 6th Framework Programme Grant (contract LSHG-CT-2003- 503496, ZF-MODELS). M. K. R. was supported by the van der Leeuw funds.

REFERENCES

Akam, M. 1987. The molecular-basis for metameric pattern in the *Drosophila* embryo. *Development* 101: 1–22.

Amores, A., et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282: 1711–1714.

Amores, A., et al. 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res*. 14: 1–10.

Aparicio, S. 2000. Vertebrate evolution—recent perspectives from fish. *Trends Genet*. 16: 54–56.

Bateson, W. 1894. *Materials for the Study of Variation Treated with Special Regard to Discontinuity in the Origin of Species*. Johns Hopkins University Press, Baltimore.

Bedford, M., Arman, E., Orr-Urtreger, A., and Lonai, P. 1995. Analysis of the Hoxd-3 gene: structure and localization of its sense and natural antisense transcripts. *DNA Cell Biol.* 14: 295–304.

Bininda-Emonds, O. R. P. 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* 261: 156.

Bruce, A. E. E., Oates, A. C., Prince, V. E., and Ho, R. K. 2001. Additional hox clusters in the zebrafish: divergent expression patterns belie equivalent activities of duplicate hoxB5 genes. *Evol. Dev*. 3: 127–144.

Carlson, M. R. J., Komine, Y., Bryant, S. V., and Gardiner, D. M. 2001. Expression of hoxb13 and hoxc10 in developing and regenerating axolotl limbs and tails. *Dev. Biol*. 229: 396–406.

de Rosa, R., et al. 1999. Hox genes in brachiopods and priapulids and protostome evolution. *Nature* 399: 772–776.

Duboule, D. 1994. Anon. *Guidebook to the Homeobox Genes*. Oxford University Press, Oxford.

Duboule, D., and Sordino, P. 1996. From fins to limbs: towards a molecular approach to the evolution of vertebrate paired appendages. *M S-Med. Sci*. 12: 147–154.

Ferrier, D. E., Minguillon, C., Holland, P. W., and Garcia-Fernandez, J. 2000. The amphioxus hox cluster: deuterostome posterior flexibility and hox14. *Evol. Dev*. 2: 284–293.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531–1545.

GarciaFernandez, J., and Holland, P. W. H. 1994. Archetypal organization of the amphioxus hox gene-cluster. *Nature* 370: 563–566.

Geraudie, J., and Borday, B. V. 2003. Posterior hoxa genes expression during zebrafish bony fin ray development and regeneration suggests their involvement in scleroblast differentiation. *Dev. Genes Evol.* 213: 182–186.

Hadrys, T., Prince, V., Hunter, M., Baker, R., and Rinkwitz, S. 2004. Comparative genomic analysis of vertebrate Hox3 and Hox4 genes. *J. Exp. Zool. B Mol. Dev. Evol*. 302: 147–164.

Hoegg, S., Brinkmann, H., Taylor, J. S., and Meyer, A. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* 59: 190–203.

Holland, P. W., and Garcia-Fernandez, J. 1996. Hox genes and chordate evolution. *Dev. Biol*. 173: 382–395.

Hsieh-Li, H.M., et al. 1995. Hoxa 11 structure, extensive antisense transcription, and function in male and female fertility. *Development* 121: 1373–1385.

Kmita, M., Fraudeau, N., Herault, Y., and Duboule, D. 2002. Serial deletions and duplications suggest a mechanism for the collinearity of Hoxd genes in limbs. *Nature* 420: 145–150.

Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform*. 5: 150–163.

Lavoie, H., et al. 2003. Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. *Hum. Mol. Genet*. 12: 2967–2979.

Lewis, E. B. 1978. Gene complex controlling segmentation in *Drosophila*. *Nature* 276: 565–570.

Malaga-Trillo, E., and Meyer, A. 2001. Genome duplications and accelerated evolution of Hox genes and cluster architecture in teleost fishes. *Am. Zool*. 41: 676–686.

McClintock, J. M., Kheirbek, M. A., and Prince, V. E. 2002. Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. *Development* 129: 2339–2354.

Meijer, A. H., et al. 2005. Transcriptome profiling of adult zebrafish at the late stage of chronic tuberculsosis due to *Mycobacterium marinum* infection. *Mol. Immunol*. 42: 1185–1203.

Meyer, A., and Schartl, M. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol*. 11: 699–704.

Nobrega, M. A., and Pennacchio, L. A. 2004. Comparative genomic analysis as a tool for biological discovery. *J. Physiol. London* 554: 31–39.

Ohno, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

Powers, T. P., and Amemiya, C. T. 2004. Evidence for a Hox14 paralog group in vertebrates. *Curr. Biol*. 14: R183–R184.

Robinson-Rechavi, M., et al. 2001. Euteleost fish genomes are characterized by expansion of gene families. *Genome Res*. 11: 781–788.

Santini, S., Boore, J. L., and Meyer, A. 2003. Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. *Genome Res*. 13: 1111–1122.

Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18: 502–504.

Seo, H. C., et al. 2004. Hox cluster disintegration with persistent anteroposterior order of expression in Oikopleura dioica. *Nature* 431: 67–71.

Strimmer, K. and von Haeseler, A. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol*. 13: 964–969.

Sordino, P., van der Hoeven, F., and Duboule, D. 1995. Hox gene expression in teleost fins and the origin of vertebrate digits. *Nature* 375: 678–681.

Swofford, D. L. 1999. *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods). Volume 4*. Sinauer Associates, Sunderland, MA.

Taylor, J. S., Van de Peer, Y., Braasch, I., and Meyer, A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. Roy. Soc. London Ser. B-Biol. Sci*. 356: 1661–1679.

Thisse, C., Thisse, B., Schilling, T. F., and Postlethwait, J. H. 1993. Structure of the zebrafish snail1 gene and its expression in wild-type, spadetail and no tail mutant embryos. *Development* 119: 1203–1215.

Thummel, R., Li, L., Tanase, C., Sarras, M. P., and Godwin, A. R. 2004. Differences in expression pattern and function between zebrafish hoxc13 orthologs: recruitment of hoxc13b into an early embryonic role. *Dev. Biol*. 274: 318–333.

Utsch, B., Becker, K., Brock, D., Lentze, M. J., Bidlingmaier, F., and Ludwig, M. 2002. A novel stable polyalanine [poly(A)] expansion in the HOXA13 gene associated with hand-foot-genital syndrome: proper function of poly(A)-harbouring transcription factors depends on a critical repeat length? *Human Genet*. 110: 488–494.

van der Hoeven, F., Zakany, J., and Duboule, D. 1996. Gene transpositions in the HoxD complex reveal a hierarchy of regulatory controls. *Cell* 85: 1025–1035.

Venkatesh, B. 2003. Evolution and diversity of fish genomes. *Curr. Opin. Genet. Dev*. 13: 588–592.

Wagner, G. P., Amemiya, C., and Ruddle, F. 2003. Hox cluster duplications and the opportunity for evolutionary novelties. *Proc. Natl. Acad. Sci. USA* 100: 14603–14606.

Zeltser, L., Desplan, C., and Heintz, N. 1996. Hoxb-13: A new Hox gene in a distant region of the HOXB cluster maintains colinearity. *Development* 122: 2475–2484.