Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 81–100

Divergence of Regulatory Sequences in Duplicated Fish Genes

R. Van Hellemont[a], *T. Blomme*[b], *Y. Van de Peer*[b], *K. Marchal*[a,c]

[a]BIOI@SCD, Dept. Electrical Engineering, K.U.Leuven, Heverlee, Leuven, [b]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Ghent, [c]Dept. Microbial and Molecular Systems, K.U.Leuven, Heverlee, Leuven, Belgium

Abstract

Duplicated genes can undergo different fates, from nonfunctionalization to subfunctionalization and neofunctionalization. In particular, changes in regulatory sequences affecting the expression domain of genes seem to be responsible for the latter two fates. In this study we used in silico motif detection to show how alterations in the composition of regulatory motifs between paralogous genes in zebrafish and *Tetraodon* might reflect the functional divergence of duplicates.

When a gene gets duplicated, it awaits four possible fates. The most likely fate is pseudogenization or nonfunctionalization [1–3]. In rare cases, one of the two duplicates acquires a new function (neofunctionalization; [4]). Subfunctionalization, where both gene copies divide the gene's original functions, forms a third potential fate [5]. Furthermore, recent studies revealed that subfunctionalization is often accompanied by neofunctionalization, which has led to a new model of gene function evolution called sub-neofunctionalization [6]. Finally, both copies can be retained, but, instead of diverging in function, they remain largely redundant and provide the organism with increased genetic robustness against harmful mutations [7, 8]. In addition, retention and redundancy of genes, at least for certain functional classes, is predicted by the 'gene balance' hypothesis, which states that retention of genes with strong dosage effects, such as for instance transcription factors, will be selected against if they are copied without their interacting partners [3, 9, 10].

In particular the subfunctionalization model [2, 5] received much attention of late, since it can, at least partially, explain the large number of genes retained after duplication events, and their subsequent functional divergence [11]. The subfunctionalization model assumes that besides depending on its protein function, the functionality of a gene is also determined by its expression domain (where and when the gene is expressed). The specific expression domain of a gene results at least partially from its transcriptional regulation, which is, in turn, encoded by a specific combination of transcription factor binding sites (defined as a regulatory module) in the gene's promoter. Each transcription factor binding site (TFBS) in a module corresponds to a DNA consensus sequence or motif that is recognized by its cognate regulatory protein or transcription factor. Changes in these TFBS can therefore be an important antecedent for expression divergence and thus for sub- or neofunctionalization [5, 8, 12–18]. However, the number of studies that show how expression divergence between paralogs is reflected by differences between regulatory elements of paralogous gene pairs is still limited [19, 20].

As a result of a genome wide fish-specific duplication event that occurred some 350 mya [21–23] and of more recent duplication events [9], ray-finned fish such as *Tetraodon* and zebrafish contain a large number of duplicated genes [24–27], of which several have already been shown to have undergone subfunctionalization [25, 28–33]. In this study, we further investigate to what extent 'in silico' analyses support expression divergence of genes through identified changes in regulatory sequences. To this end, we explicitly searched for motifs that have been preserved over 450 mya of vertebrate evolution (from mammals to ray-finned fish), but have been differentially retained in either one of two duplicates in zebrafish or *Tetraodon* and that thus might explain the experimentally observed expression divergence.

Methodology

Identification of Suitable Data Sets

In this study we focused on duplicated fish genes for which there was (some) experimental evidence that supported subfunctionalization, such as for *bmp2* [34], *glyRα* [35], *msx2* [36], *pax6* [37, 38], and *shh* [39]. In addition, several other duplicated fish genes were included, namely *efna1, en2* [5, 40], *kcnip1, ntng1, ntng2* and *six4*, and a *glyRα*-related gene family. Phylogenetic trees were constructed based on the predicted protein sequences (Ensembl [41] release 37) from human, mouse, *Tetraodon nigroviridis*, zebrafish (*Danio rerio*), rat (*Rattus norvegicus*), chicken (*Gallus gallus*), and frog (*Xenopus tropicalis*). If splice variants were reported, the longest transcript was used.

To delineate vertebrate gene families, a similarity search was performed (BLASTP, [42]; E-value cutoff E-10) with all proteins from the organisms listed above, plus those of *Ciona intestinalis* [43], version 1 and *Drosophila melanogaster* (Ensembl [41]), version 3, which were added as outgroup species. Blast hits between vertebrate sequences with a score better than the best score between a vertebrate sequence and an outgroup sequence (*Drosophila* or *Ciona*) were retained and considered members of the same gene family. The *Drosophila* or *Ciona* sequence was used to root the phylogenetic tree (see further). For each gene family, a multiple sequence alignment was created with T-Coffee 1.37 using default parameters [44]. Alignment columns containing gaps were removed when a gap was present in more than 10% of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right from the gap were also removed until a column in the sequence alignment was found where the residues were conserved in all genes included in our analyses. This was determined as follows: for every pair of residues in the column, the BLOSUM62 value was retrieved. If at least half of the pairs had a BLOSUM62 value ≥ 0, the column was considered as conserved. Neighbor joining trees (with 500 bootstrap replicates) were constructed with PHYLIP 3.5 [45] using both nucleic and amino acid sequence alignments and simple Poisson-corrected substitution models.

For all datasets, the phylogenetic tree showed genes duplicated in at least one of the fish species. These duplicates, and their homologs in human, chicken and frog were selected for further analysis (table 1). Intergenic regions of these homologs were retrieved using the Ensembl mart database release 37. Intergenic regions were defined as the region upstream of the transcription start (as defined by Ensembl), limited to 2 kb and including the 5′UTR.

Pairwise Alignment of Paralogous Intergenic Sequences

Pairwise alignments of paralogous intergenic regions were obtained with Smith-Waterman using the default parameters (gap open penalty 10 and gap extension penalty 0.5) [46]. The expected percentage identity between two unrelated intergenic sequences of the same organism was estimated by averaging the scores obtained by aligning each intergenic sequence against all other intergenic sequences of the same organism but not belonging to the same protein family.

Search for Regulatory Motifs Conserved in Each of the Gene Sets

For each gene set (i.e., all genes belonging to the same gene family), intergenic sequences were subjected to BlockSampler [47]. BlockSampler was run using default parameters – searching plus strand only ($s = 0$) and searching for

Table 1. Description of the datasets

Dataset	Newick tree[a]	Ensembl Gene IDs[a]	Experimental evidence
*bmp2**	((Xt, (Hs, Gg)), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000013409, (Dr2) ENSDARG00000041430, (Gg) ENSGALG00000008830, (Hs) ENSG00000125845, (Tn) GSTENG00020275001, (Xt) ENSXETG00000005519	RT-PCR + in situ hybridization [34]
Efna1	(Hs, ((Dr1, Tn1), (Dr2, Tn2)));	(Dr1) ENSDARG00000030326, (Dr2) ENSDARG00000018787, (Hs) ENSG00000169242, (Tn1) GSTENG00032578001, (Tn2) GSTENG00033951001	/
en2	((Xt, Hs), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000026599, (Dr2) ENSDARG00000038868, (Hs) ENSG00000164778, (Tn) GSTENG00023985001, (Xt) ENSXETG00000013496	/
*glyRα1**	(Gg, ((Dr, Tn1), Tn2)));	(Dr) ENSDARG00000006865, (Gg) ENSGALG00000004936, (Tn1) GSTENG00029286001, (Tn2) GSTENG00022245001	in situ hybridization [35]
glyRα1-related	((Xt, Gg), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000012019, (Dr2) ENSDARG00000011066, (Gg) ENSGALG00000004134, (Tn) GSTENG00024269001, (Xt) ENSXETG00000001966	/
kcnip1	((Xt, (Gg, Hs)), ((Dr1, Tn1), (Dr2, Tn2)));	(Dr1) ENSDARG00000034808, (Dr2) ENSDARG00000022109, (Gg) ENSGALG00000002132, (Hs) ENSG00000182132, (Tn1) GSTENG00020358001, (Tn2) GSTENG00024581001, (Xt) ENSXETG00000018293	/
*msx2**	((Xt, (Gg, Hs)), (Dr1, Dr2));	(Dr1) ENSDARG00000009936, (Dr2) ENSDARG00000006982, (Gg) ENSGALG00000002947, (Hs) ENSG00000120149, (Xt) ENSXETG00000009168	in situ hybridization [36]
ntng1	((Gg, Hs), (Tn1, (Dr, Tn2)));	(Dr) ENSDARG00000014973, (Gg) ENSGALG00000001896, (Hs) ENSG00000162631, (Tn1) GSTENG00027711001, (Tn2) GSTENG00035109001	/

ntng2	((Hs, Gg), ((Dr, Tn1), Tn2));	(Dr) ENSDARG00000036938, (Gg) ENSGALG00000003677, (Hs) ENSG00000196358, (Tn1) GSTENG00004089001, (Tn2) GSTENG00014392001	/
*pax6**	(((Gg, Hs), Xt), ((Dr1, Dr2), Tn));	(Dr1) ENSDARG00000045045, (Dr2) ENSDARG00000045936, (Gg) ENSGALG00000012123, (Hs) ENSG00000007372, (Tn) GSTENG00025814001, (Xt) ENSXETG00000008175	in situ hybridization + transient transfection assays + western blot analysis [37]
*shh**	((Hs, Gg), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000038867, (Dr2) ENSDARG00000039710, (Gg) ENSGALG00000006379, (Hs) ENSG00000164690, (Tn) GSTENG00023991001	in situ hybridization [39]
six4	((Xt, Hs), ((Dr1, Tn), Dr2));	(Dr1) ENSDARG00000031983, (Dr2) ENSDARG00000004695, (Hs) ENSG00000100625, (Tn) GSTENG00032223001, (Xt) ENSXETG00000016941	/

Dataset: Indicates the name of the gene family (derived from the human ortholog in the dataset). For gene sets indicated with an asterisk, experimental evidence supporting expression divergence between the fish paralogs exists. Newick tree: for each dataset the phylogenetic relations are given in Newick format. Ensembl Gene IDs: lists the genes present in each dataset by their ensembl gene ID. Experimental evidence: indicates the type of experimental evidence that supports expression divergence.

[a]Dr: *Danio rerio*, Gg: *Gallus gallus*, Hs: *Homo sapiens*, Tn: *Tetraodon nigroviridis*, Xt: *Xenopus tropicalis.*

one motif per run, prior set to 0.2, initial motif length of 8 nt, and a threshold on consensus score of 1.0. BlockSampler requires the definition of a root sequence, i.e. only conserved motifs, which are also present in the root, will be retained. As for our application the biological meaning of a root was less clear, each sequence of the gene set was chosen once as root. Per root sequence BlockSampler was run 100 times implying that the total number of runs and retrieved motifs for a gene set equaled 100 times the 'number of sequences in the gene set'. The motifs with a consensus score above 1 were selected and motifs overlapping for more than 80% were merged to avoid redundancy.

In order to account for the fact that short motifs are more likely to have a higher degree of conservation than long motifs, the consensus score of each detected block was normalized for the length of the motif using the following formula, $Cs_{ad} = (L/(L + E))\ Cs$, where L is the length of the conserved block, E is an empirical factor (set to 5) and Cs the consensus score [47].

Assessing the Statistical Significance of Detected Motifs

For each gene set, 30 random sets were compiled. These random sets have a composition similar to the genuine gene set in sequence number and origin (species), but in contrast to the genuine gene set sequences were selected randomly and as a result do not share any homology relation. For each random set, we performed the same analysis as for the genuine gene sets: BlockSampler was applied to identify conserved motifs. Per random set, the number of runs equaled 100 times the number of sequences in the random set (of which each one served once as root). After normalizing the scores, from the 100 runs of a single root the best scoring motif (highest Cs_{ad}) was selected. This resulted, for each genuine gene set in a number of random motifs equaling 30 (i.e., the number of random sets) times the 'number of sequences in this random set (i.e., number of root sequences)'. The scores of these motifs were used to estimate a random motif score distribution. To identify significant motifs in the genuine dataset we chose the Cs_{ad} of the xth percentile of this random distribution as a threshold. As a result, motifs in the genuine dataset with a Cs_{ad} higher than the chosen threshold were considered statistically significant.

Identifying Motifs Supporting Subfunctionalization

Motifs that potentially support the subfunctionalization model were identified using the following criteria: a motif was considered if it was conserved over a region of at least 8 nt and lost in at least one paralog of the fish species for which multiple paralogs were present in the gene set. In order to minimize false positive motifs, extra constraints were set on the number of additional species in which the motif had to be conserved. Indeed, if conserved over larger

Table 2. Motifs indicative for subfunctionalization with a Cs_{ad} score exceeding the 90th percentile of the random score distribution

Dataset	#	L	PI	Conservation profile	Subf sp.	Duplication type	Motif name
*bmp2**	1	29	99.5	Dr2_Gg_Hs_Tn	Dr	FSD	bmp2_1
*pax6**	4	70	90	Dr2_Gg_Hs_Tn_Xt	Dr	ZS	pax6_1
		38	90	Dr2_Gg_Hs_Tn_Xt	Dr	ZS	pax6_2
		71	90	Dr2_Gg_Hs_Tn_Xt	Dr	ZS	pax6_3
		50	90	Dr2_Gg_Hs_Tn_Xt	Dr	ZS	pax6_4
*shh**	2	15	95	Dr2_Gg_Hs_Tn	Dr	FSD	shh_1
		16	99	Dr2_Gg_Hs_Tn	Dr	FSD	shh_2
kcnip1	1	13	95	Dr1_Dr2_Gg_Hs_Tn1_Xt	Tn	FSD	Kcnip1_1
Total	8						

Dataset: The gene set in which the motifs were detected. Gene sets indicated with an asterisk contain fish paralogs for which subfunctionalization has been shown in literature. #: the number of motifs detected in the gene set that support subfunctionalization given this threshold. L: the length of the motif indicative for subfunctionalization (number of nucleotides) PI: indicates the percentile of the random distribution to which the score of the motif belongs. Conservation profile: indicates in which homologs of the gene family the motif was also present (see footnote of table 1). Subf sp.: indicates in which fish-species the motif was lost. Duplication type: indicates from which duplication event the paralogs originated for which a motif was found that supports subfunctionalization (FSD: ancient fish specific duplication event; TS: *Tetraodon* specific duplication event; ZS: zebrafish specific duplication event). Motif name: the name to unambiguously indicate a specific motif.

phylogenetic distances, we can be more confident in the motif prediction. These constraints depended on the composition of the gene set: If the gene set under study consisted of multiple non-fish homologs, either frog, chicken or human (which was the case for *bmp2*, *en2*, *glyRα*-related, *kcnip1*, *msx2*, *ntng1*, *ntng2*, *pax6*, *shh*, *six4*), a motif was only considered if it was conserved in at least two non-fish species. If the motif under study was derived from a gene set that contained only one non-fish homolog (*efna1* and *glyRα*), the motif had to be present in this one non-fish sequence.

For each motif we constructed a profile that indicates whether or not the motif occurs in the respective species from which the homologs of the gene family were derived. If the profile of the motif satisfies the requirements mentioned above, its profile was said to support subfunctionalization. Phylogenetic profiles of motifs supporting subfunctionalization are represented in tables 2 and 3. To assess whether the detected motifs correspond to known transcription factor binding sites, we scanned the human instance of each conserved motif with the Transfac 8.2 database of vertebrate transcription factor binding site profiles [48]. This scanning was performed using MotifLocator [49, 50] with a

Table 3. Motifs indicative for subfunctionalization for the gene sets for which a relaxed selection criterium was used

Dataset	#	L	PI	Conservation profile	Subf sp.	Duplication type	Motif name
efna1	2	12	75	Dr1_Dr2_Hs_Tn2	Tn	TS	efna1_1
		8	55	Dr1_Dr2_Hs_Tn2	Tn	TS	efna1_2
Total	2						

For column descriptions and legend see tables 1 and 2.

0th order vertebrate background model. Hits with a score >0.9 were regarded as potential binding sites. The binding sites are indicated by the Transfac factor name [48]. To further validate the link between the binding sites revealed with this screening and the gene under study, we did a text-based search with PubMed [51] using the name of the gene/protein under study (e.g. *pax6*) and the name of the transcription factor potentially binding the promoter region of this gene as search terms. When such a link existed, this is explicitly mentioned in the results section.

Results

The goal of our study was to see whether divergent expression of duplicated genes is reflected in any detectable way by a different composition of their regulatory sequences, in particular by the presence or absence of specific motifs. First, this requires identifying interesting case studies, i.e., gene families that contain members of fish specific duplication events. Second, we need to compile the potential regulatory motifs present in the intergenic sequences of these gene families and to identify which of the motifs have been differentially retained in one of the fish paralogs. However, since the regulatory motifs present in fish genomes are still largely unknown, the list of potential motifs was compiled based on comparative de novo motif detection methods, better known as phylogenetic footprinting. Phylogenetic footprinting assumes that biologically relevant sequences, such as regulatory motifs, evolve slower than their surrounding non-functional intergenic sequences. By using cross-species conservation, short stretches of DNA that are conserved over certain phylogenetic distances are identified as potential motifs. The greater the phylogenetic distance over which the motif is conserved and the more orthologs in which the

motif can be detected are present, the more confidence can be put in this prediction.

However, as we are specifically searching for conserved motifs that are differentially lost between paralogs, we had to rely on a phylogenetic footprinting methodology that is able to align strongly evolved sequences of which some do not contain the motifs [47].

Identifying Gene Sets Containing Duplicated Fish Genes

Our analysis was performed on a selection of gene families that contained paralogs either originating from a duplication event before the divergence of zebrafish and *Tetraodon* (further referred to as the ancient fish specific duplication, FSD) or from a more recent duplication specific to either zebrafish or *Tetraodon* (that occurred after divergence of both species [9]) (for a complete list of these gene families see table 1). For some of these fish-specific paralogs, subfunctionalization was supported by literature (e.g., *bmp2* [34, 39], *glyRα* [35], *msx2* [36], *pax6* [37, 38] and *ssh* [39] (see table 1)).

For gene sets *bmp2*, *efna1*, *en2*, *glyRα1*, *glyRα1*-related, *kcnip1*, *ntng1*, *ntng2*, *pax6*, *shh* and *six4*, the topology of the corresponding phylogenetic trees indicates that the paralogs resulted from the ancient FSD event which took place before the divergence of zebrafish and *Tetraodon* (about 150 mya [24]). For the *pax6* gene family, the two zebrafish copies are the result of a more recent zebrafish specific duplication event (table 1). Concerning the *msx2* gene family, the topology did not allow us to conclude whether the *msx2* zebrafish copies resulted from the ancient FSD or whether they were the result of a more recent duplication event in zebrafish.

Determining the Overall Homology Between Paralogous Intergenic Regions

In order to determine their overall conservation, intergenic paralogous regions in fish were aligned using Smith-Waterman [46]. Results are shown in table 4. The conservation level of paralogous intergenic regions resulting from the ancient FSD [9] (*Tetraodon*: 40.4% and zebrafish: 43.6%) was comparable to that of unrelated sequences, which was estimated 40.4% and 43% for *Tetraodon* and *Danio rerio* respectively (see methods section). This analysis also indicates that in these ancient duplicates, except for the conserved regulatory motifs, no sequence conservation is to be expected.

Identification of Motifs Supporting Subfunctionalization

To search for differences in motif composition between duplicates, we first compiled all potential motifs conserved within the intergenic regions of genes belonging to the same gene family (see material and methods). To this end we

Table 4. Intergenic homology between fish duplicates

Dataset	Compared duplicates	Type	Identity (%)
*bmp2**	Dr1–Dr2	FSD	44.4
*glra**	Tn1–Tn2	FSD	41.4
*msx2**	Dr1–Dr2	FSD/ZS	44.5
*pax6**	Dr1–Dr2	ZS	39.2
*shh**	Dr1–Dr2	FSD	43.6
efna1	Dr1–Dr2	FSD	43.9
	Tn1–Tn2	FSD	32.6
en2	Dr1–Dr2	FSD	45.9
glra-related	Dr1–Dr2	FSD	42.3
kcnip1	Dr1–Dr2	FSD	43.4
	Tn1–Tn2	FSD	41.9
ntng1	Tn1–Tn2	FSD	42.3
ntng2	Tn1–Tn2	FSD	43.9
six4	Dr1–Dr2	FSD	41.7

Intergenic regions of fish paralogs present in the gene sets under study were pairwise aligned using Smith-Waterman [46]. Dataset: indicates the name of the gene family (derived from the human ortholog in the dataset) from which the fish paralogs were compared. For gene sets indicated with an asterisk, experimental evidence supporting expression divergence between the fish paralogs exists. Compared duplicates: the fish genes for which the intergenic sequences were aligned; for the corresponding Ensembl gene IDs and legend we refer to table 1. Type: the type of duplication event the duplicates are the result of (based on the phylogenetic trees); FSD: ancient fish specific duplication; ZS: zebrafish specific duplication. Percent identity: the similarity between the intergenic sequences of the aligned duplicates.

used BlockSampler, a procedure based on Gibbs sampling that searches for statistically overrepresented motifs. In the presence of an appropriate background model, the procedure is known to be quite robust against noise, i.e., sequences that do not contain the motif [52–54]. In the context of subfunctionalization, this property is essential as it allows finding motifs that are not conserved in all branches of the phylogenetic tree. For each set, applying BlockSampler resulted in a list of conserved motifs. To detect motifs supporting subfunctionalization, from this list we selected those motifs that were significantly conserved but missing in at least one of the fish paralogs, either *Tetraodon* or zebrafish (i.e., the species for which multiple paralogs are present in the gene set under study). Especially for the ancient duplications for which the overall similarity in intergenic sequences between the fish paralogs is quite low, many differences are

expected to be found in their promoter regions, most of which probably do not correspond to biologically relevant subfunctionalized motifs. Therefore, in order to select the most relevant predictions we considered only those motifs that were also conserved in phylogenetic lineages other than fish and thus were preserved over 450 mya of vertebrate evolution (see materials and methods for the exact criteria).

To test to what extent the choice of the threshold on the motif scores (defined as the xth percentile of the random scores) determined the total number of motifs retrieved and thus the number of gene sets for which we detected a motif(s) indicative for subfunctionalization, we repeated the analysis for multiple threshold levels (ranging from the 99.5th to the 50th percentile of the random score distribution). The results for gene sets containing homologs from multiple non-fish species are summarized in table 2, considering the 99.5th, 99th, 95th and 90th percentile of the random distribution as threshold on the motif scores. As expected, lowering the threshold of our search allows detecting more motifs indicative for subfunctionalization. However, as the stringency of the search becomes lower, the motifs taken into account become gradually shorter and presumably less reliable.

When using a quite conservative threshold (motif scores exceeding the 90th percentile of the random distribution), in three (*bmp2*, *pax6* and *shh*) out of the five datasets for which expression divergence was experimentally demonstrated, we could find at least one motif indicative for subfunctionalization. Besides in these experimentally supported datasets, we also found motif-based indications for subfunctionalization in *efna1* (although only when using a relaxed threshold in the motif scores, table 3) and *kcnip1*.

Detailed Description of the Datasets with Subfunctionalized Motifs

Figures 1 to 5 display the results for the datasets *bmp2*, *pax6*, *shh*, *kcnip1* and *efna1*. Significantly overrepresented motifs are mapped. An arrow indicates motifs that might be supportive of subfunctionalization. Below we give a more detailed description of these results.

In vertebrates, bone morphogenetic proteins **(Bmps)** play a crucial role in establishing the early body plan and in organogenesis [55]. Martinez-Barbera et al. [34] studied the expression pattern of zebrafish *bmp2* paralogs, *bmp2a* and *bmp2b*. They found indications for divergent expression profiles in the gastrulating embryo and in the pectoral fin bud. In this study, the *bmp2* gene family consists of two zebrafish genes (table 1, fig. 1) that correspond to the genes studied by Martinez-Barbera et al. [34]. The motif indicated in figure 1 (table 2) that has been retained in one zebrafish copy (Dr2, ENSDARG00000041430) but that was lost in the other (Dr1, ENSDARG00000013409), could possibly explain this observed divergence.

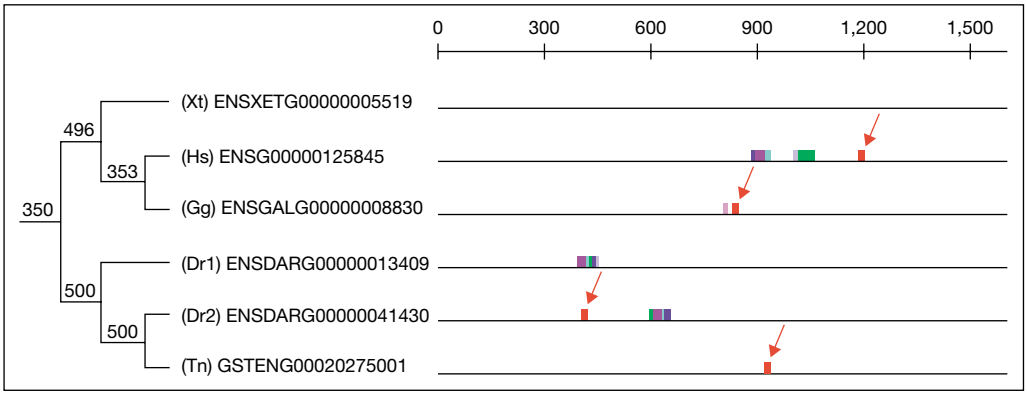

Fig. 1. Graphical display of the motifs found in the upstream regions of *bmp2*. Graphical display of all motifs with a score exceeding the 90th percentile of the random score distribution. The motifs that are in support of subfunctionalization are indicated by an arrow. The phylogenetic tree (branch lengths not drawn to scale) illustrates the evolutionary relationships between the homologs; these are indicated as defined in table 1. Abbreviations used: Xt: *Xenopus tropicalis*, Hs: *Homo sapiens*, Gg: *Gallus gallus*, Dr: *Danio rerio*, Tn: *Tetraodon nigroviridis.*

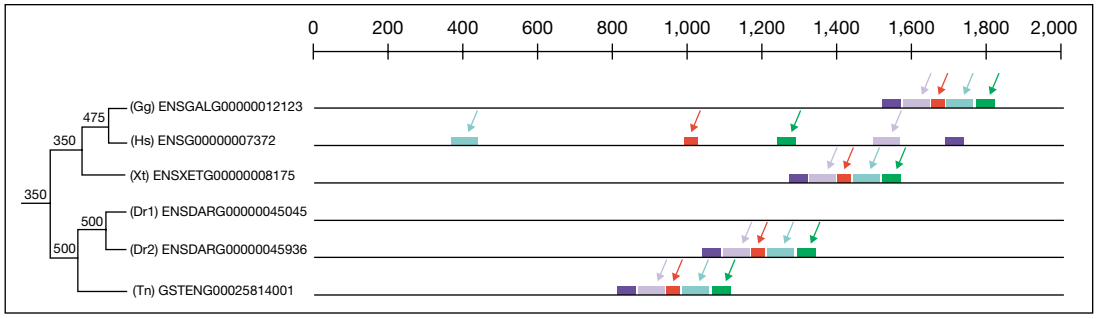

Fig. 2. Graphical display of the motifs found in the upstream regions of *pax6.* Interpretation is as in figure 1.

Pax6 plays an important role in the central nervous system and in the developing eye of both vertebrates and invertebrates. According to our analysis, the *pax6* gene family contains two zebrafish paralogs, which (given the position of the *Tetraodon* homolog in the tree topology) originated from a zebrafish specific duplication (fig. 2). The presence of two zebrafish paralogs is consistent with the observations of Nornes et al. [37]. They observed that both zebrafish

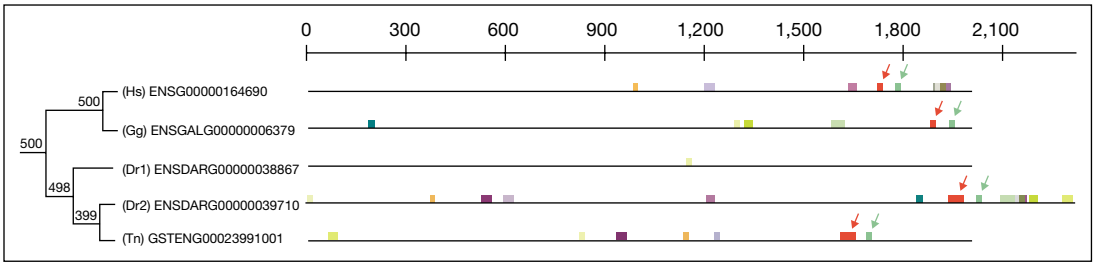

Fig. 3. Graphical display of the motifs found in the upstream regions of *shh.* Interpretation is as in figure 1.

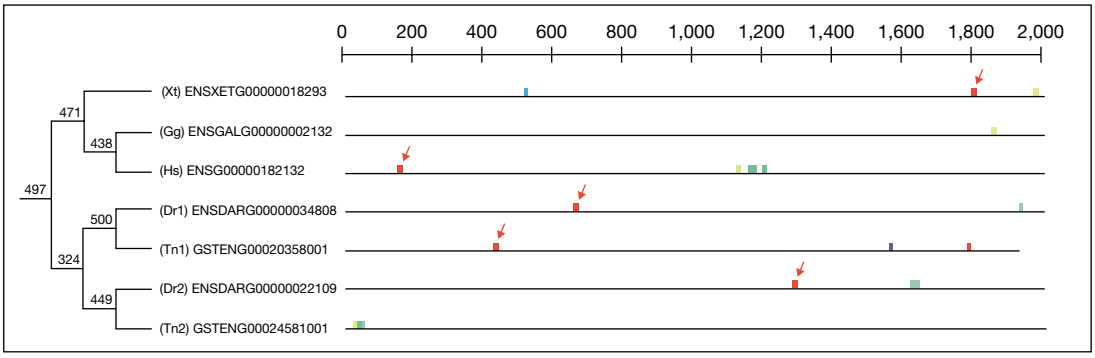

Fig. 4. Graphical display of the motifs found in the upstream regions of *kcnip1.* Interpretation is as in figure 1.

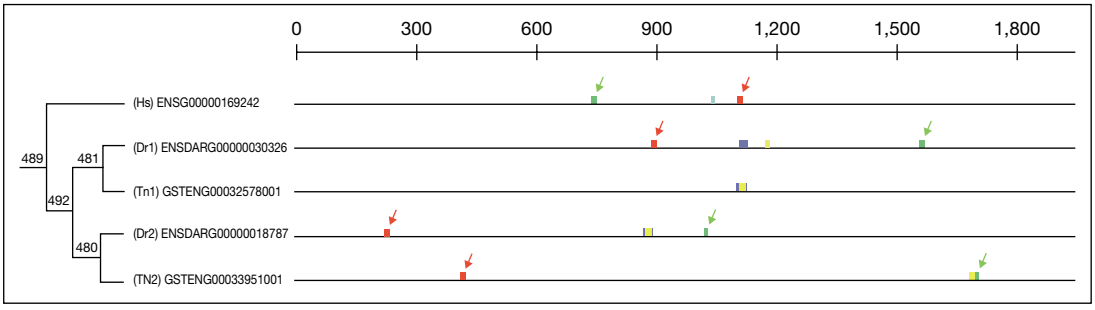

Fig. 5. Graphical display of the motifs found in the upstream regions of *efna1.* Interpretation is as in figure 1.

copies have unique expression domains that sum up to the total expression domain for the single *pax6* copy present in birds and mammals [2]. Figure 2 displays the conserved motifs identified in the promoter region of the *pax6* homologs. Motifs that might be indicative for subfunctionalization are indicated by arrows: we identified four motifs conserved in human, chicken, frog, *Tetraodon* and in one zebrafish paralog (Dr2, ENSDARG00000045936; table 2). The complete absence of all of these motifs in the zebrafish paralog (Dr1, ENSDARG00000045045) can also be interpreted as an indication for nonfunctionalization of this paralog. However, because experimental evidence about the expression of both zebrafish paralogs exists [37], subfunctionalization seems the more likely fate. The order in which the motifs occur in the intergenic regions seems to be perfectly conserved in the non-mammalian sequences where they are concatenated into a large conserved region of circa 250 nt (fig. 2). In the human ortholog on the contrary, the order and spacing of these motifs seems to be altered. In order to get an idea of the binding sites localized in the four motifs reported here, we screened the human motif instance with the Transfac database of transcription factor binding sites. As is summarized in table 5, different potential binding sites are present in the *pax6* motifs. For instance, pax6_3 contains an AP-2α and an AP-2rep binding site. This is plausible, since both Pax6 and AP-2α function in eye development [56]. Moreover, both transcription factors are known to interact in coordinating corneal epithelial repair [57].

Vertebrate **hedgehog** genes are involved in many developmental processes [58]. As Laforest et al. showed that zebrafish hedgehog paralogs exhibit expression patterns that suggest subfunctionalization, we chose to study the *sonic hedgehog* or *shh* gene family [39] in more detail. Figure 3 illustrates two significant motifs (see arrows) that possibly support subfunctionalization (see also table 2). These motifs, indicated in red and green, have both been conserved in human, chicken, *Tetraodon* and Dr2 (ENSDARG00000039710) but were lost in Dr1 (ENSDARG00000038867). The order and spacing between these two motifs also seems to have been retained during evolution. Besides these, figure 3 also displays some additional interesting motifs pointing towards subfunctionalization (for instance, the dark purple and dark yellow motifs). These were not initially retained as 'significant motifs' under our strong selection criteria, because they were either too short or not conserved in multiple non-fish species.

kcnip1 encodes the potassium channel-interacting protein [59]. In this study we identified a frog, chicken and human homolog, two zebrafish and two *Tetraodon* paralogs. The tree topology (fig. 4) indicates that the four fish genes are the result of an ancient FSD. As is shown in figure 4, we identified one motif that is in support of a possible divergent expression profile (table 2). This

Table 5. The potential transcription factor binding sites located in the detected motifs indicative for subfunctionalization

Motif Name	Consensus and possible binding sites
bmp2_1	TTGTTTTGTTTTGTTTTTT SRY, M00148: AAACWAM: 5-11 – (1.0); 10-16 – (1.0)
pax6_1	GGCTCGAGGGCCAGGTTGAGGGTACTCATCGAGCCTCGAACTCCTCCTAAAAATGATTCCTGCCAAAAGC Cap, M00253, NCANHNNN: 49-56 – (0.963) CdxA, M00101, AWTWMTR: 46-52 – (0.904) Hnf4, M00967, AARGTCCAN: 6-14 + (0.931) Etf, M00695, GVGGMGG: 43-49 – (0.906) Lyf-1, M00141, TTTGGGAGR: 59-67 – (0.931); 43-51 – (0.950) NF1, M00193, NNTTGGCNNNNNNCCNNN: 51-68 – (0.919)
pax6_2	ACCACTGTCACTTTCAAATTGGAGAGCCAGATGGAAGC E2a, M00804, GGCGSG: 21-34 – (0.907) Irf, M00772, BNCRSTTTCANTTYY: 1-15 + (0.946) Tal1, M00993, TCCAKCTGNY: 26-35 – (1.0)
pax6_3	TGGTAAGGTCTAGGCCCAGACTAGAGTGGCCAGTGGGAGGTGGGCGCTCCTAGGCCTTAACACAGGATGCC AP-2α, M00469, GCCNNNRGS: 29-37 + (0.917) AP-2rep, M00468, CAGTGGG: 31-37 + (1.0) Cap, M00253, NCANHNNN: 16-23 + (0.906); 63-70 – (0.925); 36-43 – (0.904) CCAAT box, M00254, NNNRRCCAATSA: 25-36 + (0.933) C/EBP, M00159, NNTKTGGWNANNN: 54-66 – (0.906) CHCH, M00986, CGGGNN: 34-39 + (0.932) Etf, M00695, GVGGMGG: 34-40 + (0.911) Ets, M00971, ACTTCCTS: 63-70 – (0.925) Pea3, M00655, ACWTCCK: 64-70 – (0.933)

Table 5. (continued)

Motif Name	Consensus and possible binding sites
pax6_4	ATTTTCCTGTTTTCCTCCTCTAAGTCACAAAGTCAACAGTTAATTCAAAG AP-1, M00172, RSTGACTNMNW: 19-29 – (0.942) AP-1, M00517, NNNTGAGTCAKCN: 18-30 – (0.905) AP-1, M00924, TGACTCANNSKN: 16-27 – (0.901) AP-1, M00926, TGAGTCAN: 21-28 + (0.904) Fox, M00809, KATTGTTTRTTTW: 29-41 – (0.953) Hnf3α, M00724, TRTTTGYTYWN: 28-38 – (0.932) Hnf3β, M00131, KGNANTRTTTRYTTW: 29-43 – (0.908) Pou1f1, M00744, ATGAATAAWT: 39-48 – (0.915) Sf1, M00727, TGRCCTTG: 28-35 – (0.918) Stat1, M00496, NNTTTCCN: 1-8 + (0.948); 9-16 + (0.9704) Stat6, M00500, NNYTTCCY: 9-16 + (0.915)
shh_1	GCTCTCCAGGCTTGC
shh_2	TCAGATGCGCCCCTGG
kcnip1_1	TGTGTATCTGTGT
efna1_1	ACGCAGACACACA
efna1_2	ATGTTTATT

Motif name: The name to unambiguously indicate a certain motif detected (with a Cs_{ad} score exceeding 90th percentile of the random score distribution for *bmp2*, *pax6*, *ssh* and *kcnip1* and a Cs_{ad} score exceeding 50th percentile for *efna1*). These names correspond to the ones in tables 2 and 3. Consensus and possible binding sites: the sequence of the motif in the intergenic region of the human homolog (table 1) is given followed by the possible binding sites situated in this motif (Transfac name, Transfac ID, consensus sequence, positions, strand and score). Remark: For the shortest motifs MotifLocator could not be used to screen for potential binding sites. Therefore only the motif instance in human is given.

motif, indicated in red, is retained in frog, human, both zebrafish paralogs, and one *Tetraodon* **paralog** (Tn1, GSTENG00020358001). Two other interesting motifs (the green and light blue motifs) were detected in Tn2, GSTENG00 024581001: both these motifs seem to be present in the human sequence but are divergently retained over the fish paralogs. The smaller blue motif is retained in Dr1 and Tn1, while the green motif is retained in Dr2 and Tn2. From the tree topology it seems that the combined motif, still present in the human sequence might have been subfunctionalized after an early fish duplication that took place before the speciation between *Tetraodon* and zebrafish. The motifs seem to have a classical pattern of subfunctionalization. Note, however, that we did not primarily retain them as they do not meet our selection criteria (the motifs are conserved in one non-fish homolog only).

Also in ***efna1***, which encodes an ephrin-A1 precursor, we found two motifs indicative of expression divergence between paralog Tn1 (GSTEN G00032578001) and paralog Tn2 (GSTENG00033951001; fig. 5). These two motifs, respectively 12 and 8 bp long, and conserved in the intergenic sequences of the human homolog were also present in both zebrafish paralogs, but only in one of the two *Tetraodon* paralogs (Tn2; see table 3).

Discussion

In this study, we found indications that expression divergence between paralogs in zebrafish and/or *Tetraodon* is reflected by differences in regulatory motifs. We investigated five gene families for which experimental evidence supported subfunctionalization. For three of these proof-of-concept gene families, we identified at least one motif that was differentially lost after a fish-specific duplication event and that seemed to be in accordance with the experimentally observed expression divergence. Besides in the 'proof-of-concept' datasets, we found differential alterations in regulatory motifs between the fish paralogs that point towards potential subfunctionalization in two other gene families (*efna1* and *kcnip1*).

In order to assess which potential transcription factors bind to the conserved motifs, we screened them with the Transfac database of TFBS. Several potential TFBS seemed to be present in the conserved motifs but to our knowledge, for the majority of these TFBS no clear link with the genes containing the conserved motifs was found in literature.

The sequence dependent indications for subfunctionalization identified in this study of course largely depend on the reliability of our in silico predicted regulatory motifs. To select confident predictions, we used strict selection criteria and considered only those motifs that were conserved over at least 450 mya of vertebrate

evolution. On the other hand, by using these conservative selection criteria we probably discard many functional motifs. Indeed, motifs that are too short, too degenerated, or very lineage specific will remain undetected. As a result, we most likely underestimate the number of motifs indicative for subfunctionalization. This might explain why we only find in a subset of the 'proof-of-concept' datasets sequence based indications for subfunctionalization. Moreover, according to its strict definition, subfunctionalization implies that both paralogs divide the gene's original function over both gene copies. When relating this to expression divergence and subsequent changes in regulatory motifs, one expects to find two ancestral motifs still present in an outgroup species to be divided between the two paralogous intergenic regions. We could not detect any example of this idealized situation of subfunctionalization due to our conservative approach; however, when using more relaxed criteria our method identified such an example.

Despite our conservative strategy, in nearly half of the tested datasets clear sequence based indications for potential expression divergence were present, indicating that subfunctionalization is probably more general than is assumed at this point (see also [8]).

Acknowledgements

T. Blomme and R. Van Hellemont are fellows of the IWT. This work is partially supported by: 1. IWT projects: GBOU-SQUAD-20160; 2. Research Council KULeuven: GOA Mefisto-666, GOA-Ambiorics, EF/05/007 SymBioSys, IDO genetic networks; 3. FWO projects: G.0115.01, G.0413.03 and G.0318.05; 4. IUAP V-22 (2002–2006). We would like to thank S. Robbens for the assistance with the construction of phylogenetic trees.

References

1 Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. Science 2000;290: 1151–1155.
2 Lynch M, Force A: The probability of duplicate gene preservation by subfunctionalization. Genetics 2000;154:459–473.
3 Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, et al: Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 2005;102:5454–5459.
4 Taylor JS, Raes J: Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 2004;38:615–643.
5 Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of duplicate genes by complementary, degenerative mutations. Genetics 1999;151:1531–1545.
6 He X, Zhang J: Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 2005;169:1157–1164.
7 Gu X: Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet 2003;19:354–356.
8 Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y: Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. Genome Biol 2006;7:R13.1–R13.11.

9 Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 2006;7:R43.1–R43.12.
10 Freeling M, Thomas BC: Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 2006;16:805–814.
11 Moore RC, Purugganan MD: The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol 2005;8:122–128.
12 Papp B, Pal C, Hurst LD: Evolution of *cis*-regulatory elements in duplicated genes of yeast. Trends Genet 2003;19:417–422.
13 De Bodt S, Theissen G, Van de Peer Y: Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. Mol Biol Evol 2006;23:1293–1303.
14 Prince VE, Pickett FB: Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet 2002;3:827–837.
15 Gu X, Zhang Z, Huang W: Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci USA 2005;102:707–712.
16 Gu Z, Nicolae D, Lu HH, Li WH: Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 2002;18:609–613.
17 Zhang Z, Gu J, Gu X: How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? Trends Genet 2004;20:403–407.
18 Kafri R, Bar-Even A, Pilpel Y: Transcription control reprogramming in genetic backup circuits. Nat Genet 2005;37:295–299.
19 Chang L, Khoo B, Wong L, Tropepe V: Genomic sequence and spatiotemporal expression comparison of zebrafish *mbx1* and its paralog, *mbx2*. Dev Genes Evol 2006;216:647–654.
20 Jimenez-Delgado S, Crespo M, Permanyer J, Garcia-Fernandez J, Manzanares M: Evolutionary genomics of the recently duplicated amphioxus Hairy genes. Int J Biol Sci 2006;2:66–72.
21 Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B: Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol 2004;21:1146–1151.
22 Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al: Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 2004;431:946–957.
23 Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y: Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci USA 2004;101:1638–1643.
24 Meyer A, Van de Peer Y: From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 2005;27:937–945.
25 Postlethwait JH, Yan YL, Gates MA, Horne S, Amores A, et al: Vertebrate genome evolution and the zebrafish gene map. Nat Genet 1998;18:345–349.
26 Van de Peer Y, Taylor JS, Meyer A: Are all fishes ancient polyploids? J Struct Funct Genomics 2003;3:65–73.
27 Wittbrodt J, Meyer A, Schartl M: More genes in fish? BioEssays 1998;20:511–515.
28 Van de Peer Y, Taylor JS, Braasch I, Meyer A: The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J Mol Evol 2001;53:436–446.
29 Altschmied J, Delfgaauw J, Wilde B, Duschl J, Bouneau L, et al: Subfunctionalization of duplicate *mitf* genes associated with differential degeneration of alternative exons in fish. Genetics 2002;161:259–267.
30 Bollig F, Mehringer R, Perner B, Hartung C, Schafer M, et al: Identification and comparative expression analysis of a second *wt1* gene in zebrafish. Dev Dyn 2006;235:554–561.
31 Volff JN: Genome evolution and biodiversity in teleost fish. Heredity 2005;94:280–294.
32 Winkler C, Schafer M, Duschl J, Schartl M, Volff JN: Functional divergence of two zebrafish midkine growth factors following fish-specific gene duplication. Genome Res 2003;13:1067–1081.
33 Postlethwait J, Amores A, Cresko W, Singer A, Yan YL: Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet 2004;20:481–490.
34 Martinez-Barbera JP, Toresson H, Da Rocha S, Krauss S: Cloning and expression of three members of the zebrafish Bmp family: *Bmp2a*, *Bmp2b* and *Bmp4*. Gene 1997;198:53–59.
35 Imboden M, Devignot V, Goblet C: Phylogenetic relationships and chromosomal location of five distinct glycine receptor subunit genes in the teleost *Danio rerio*. Dev Genes 2001;211:415–422.

36 Ekker M, Akimenko MA, Allende ML, Smith R, Drouin G, et al: Relationships among msx gene structure and function in zebrafish and other vertebrates. Mol Biol Evol 1997;14:1008–1022.
37 Nornes S, Clarkson M, Mikkola I, Pedersen M, Bardsley A, et al: Zebrafish contains two *pax6* genes involved in eye development. Mech Dev 1998;77:185–196.
38 Force A, Shashikant C, Stadler P, Amemiya CT: Comparative genomics, *cis*-regulatory elements, and gene duplication. Methods Cell Biol 2004;77:545–561.
39 Laforest L, Brown CW, Poleo G, Geraudie J, et al: Involvement of the sonic hedgehog, *patched 1* and *bmp2* genes in patterning of the zebrafish dermal fin rays. Development 1998;125:4175–4184.
40 Joyner AL, Martin GR: *En-1* and *En-2*, two mouse genes with sequence homolog to the *Drosophila engrailed* gene: expression during embryogenesis. Genes Dev 1987;1:29–38.
41 Ensembl genome browser [http://www.ensembl.org]
42 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
43 JGI [http://genome.jgi-psf.org]
44 Notredame C, Higgins DG, Heringa J: T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
45 Felsenstein J: PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 1989;5:164–166.
46 Smith TF, Waterman MS: Comparison of biosequences. Adv Appl Math 1981;2:482–489.
47 Van Hellemont R, Monsieurs P, Thijs G, De Moor B, Van de Peer Y, Marchal K: A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biol 2005;6:R113.1–R113.18.
48 Wingender E, Chen X, Fricke E, Geffers R, Hehl R, et al: The TRANSFAC system on gene expression regulation. Nucleic Acids Res 2001;29:281–283.
49 Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: Toucan: deciphering the *cis*-regulatory logic of coregulated genes. Nucleic Acids Res 2003;31:1753–1764.
50 Coessens B, Thijs G, Aerts S, Marchal K, De Smet F, et al: INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. Nucleic Acids Res 2003;31:3468–3470.
51 PubMed [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed]
52 Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B, Vanderleyden J: Genome-specific higher-order background models to improve motif detection. Trends Microbiol 2003;11:61–66.
53 Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, et al: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 2001;17:1113–1122.
54 Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, et al: A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 2002;9:447–464.
55 Hogan BL: Bone morphogenetic proteins: multifunctional regulators of vertebrate development. Genes Dev 1996;10:1580–1594.
56 West-Mays JA, Zhang J, Nottoli T, Hagopian-Donaldson S, Libby D, et al: AP-2alpha transcription factor is required for early morphogenesis of the lens vesicle. Dev Biol 1999;206:46–62.
57 Sivak JM, West-Mays JA, Yee A, Williams T, Fini ME: Transcription factors Pax6 and AP-2alpha interact to coordinate corneal epithelial repair by controlling expression of matrix metalloproteinase gelatinase B. Mol Cell Biol 2004;24:245–257.
58 Ingham PW, McMahon AP: Hedgehog signaling in animal development: paradigms and principles. Genes Dev 2001;15:3059–3087.
59 Shibata R, Misonou H, Campomanes CR, Anderson AE, Schrader LA, et al: A fundamental role for KChIPs in determining the molecular properties and trafficking of Kv4.2 potassium channels. J Biol Chem 2003;278:36445–36454.

Kathleen Marchal
Department of Microbial and Molecular Systems (CMPG)
K.U.Leuven, Kasteelpark Arenberg 20
B-3001 Heverlee, Leuven (Belgium)
Tel. +32169685, Fax +3216321963, E-Mail Kathleen.marchal@biw.kuleuven.be