Development 139, 917-928 (2012) doi:10.1242/dev.071803

Cis-regulatory properties of medaka synexpression groups

Mirana Ramialison[1,2,*,‡], Robert Reinhardt[1], Thorsten Henrich[3], Beate Wittbrodt[1], Tanja Kellner[1], Camille M. Lowy[1] and Joachim Wittbrodt[1,2,‡]

SUMMARY

During embryogenesis, tissue specification is triggered by the expression of a unique combination of developmental genes and their expression in time and space is crucial for successful development. Synexpression groups are batteries of spatiotemporally co-expressed genes that act in shared biological processes through their coordinated expression. Although several synexpression groups have been described in numerous vertebrate species, the regulatory mechanisms that orchestrate their common complex expression pattern remain to be elucidated. Here we performed a pilot screen on 560 genes of the vertebrate model system medaka (*Oryzias latipes*) to systematically identify synexpression groups and investigate their regulatory properties by searching for common regulatory cues. We find that synexpression groups share DNA motifs that are arranged in various combinations into cis-regulatory modules that drive co-expression. In contrast to previous assumptions that these genes are located randomly in the genome, we discovered that genes belonging to the same synexpression group frequently occur in synexpression clusters in the genome. This work presents a first repertoire of synexpression group common signatures, a resource that will contribute to deciphering developmental gene regulatory networks.

KEY WORDS: Chromosomal clustering, Cis-regulatory modules, Collinearity, Motif discovery, Medaka, Synexpression

INTRODUCTION

Tissue determination during embryogenesis is governed by dynamic developmental gene expression, and failures of proper gene expression in time and space lead to embryonic defects. Spatiotemporal gene expression is tightly controlled within a developmental gene regulatory network (GRN) (Peter and Davidson, 2009). A major challenge in the genomic era is to decipher this network by identifying its components and the relationships between them.

Subsets of developmental genes called synexpression groups (SGs) are characterised by strikingly similar spatiotemporal gene expression patterns throughout embryo development. They have been proposed to be involved in common biological pathways, highlighting their central role as basic modules of developmental GRNs, promoting the coordinate expression of gene batteries and potentiating rapid evolutionary changes (Niehrs and Pollet, 1999). To identify SGs, a number of large-scale whole-mount gene expression screens have been performed in *Drosophila* (Tomancak et al., 2002), *Xenopus* (Gawantka et al., 1998) and mouse (Visel et al., 2007). These studies have highlighted the usefulness of SGs in predicting biological functions for uncharacterised genes, based on the assumption that genes belonging to the same SG are involved in a common biological process (Amaya, 2005).

Niehrs and Pollet (Niehrs and Pollet, 1999) have proposed that genes within an SG display a similar expression pattern because they possess common cis-regulatory elements that are bound by a common regulatory trans-acting factor (t-AF). Thus, genes belonging to the same SG should share common sequence motifs that are reflective of the t-AF DNA binding site. Support for this hypothesis has been provided for several SGs in amniotes (Ernsberger, 2000), *Ciona* (Brown et al., 2007), *Xenopus* (Karaulanov et al., 2004), medaka (Ramialison et al., 2008), mouse (Visel et al., 2007) and human (Grade et al., 2009).

Although these studies have been performed in specific SGs, they suggest that their shared cis-regulatory input leads to co-expression. In this pilot analysis we present a systematic investigation of the functional and regulatory properties of selected SGs in the developing medaka fish embryo with a view to uncovering general principles of SG regulation.

First, we identified SGs by collecting similar spatiotemporal gene expression patterns from the Medaka Expression Pattern Database (MEPD) (Henrich et al., 2003; Henrich et al., 2005) and from a new random in situ hybridisation (ISH) screen. From these data, we built a network of co-expressed genes at three developmental time points. We used data reduction to collapse highly similar gene expression patterns into expression 'mountains', followed by a higher-order hierarchical clustering of these mountains into 'mountain ranges'. From this co-expression network, SGs were identified by systematically searching for modules of co-expressed genes that share a common biological process. By looking for over-represented DNA motifs de novo in each SG, we revealed that SGs share common complex putative regulatory input consisting of different cis-regulatory modules (CRMs) that are potentially able to drive synexpression. We performed an in-depth analysis of an SG of genes expressed in proliferative tissues and demonstrated experimentally the ability of various CRMs to recapitulate individual components of the overall expression pattern of the SG. Finally, we systematically examined the chromosomal localisation of genes from each SG and discovered that some SGs contain genes that are clustered at the same genomic locus, a feature of SGs that has not been previously described.

[1]University of Heidelberg, Centre for Organismal Studies, 69120 Heidelberg, Germany. [2]Karlsruhe Institute of Technology, Institute of Toxicology and Genetics, 76344 Karlsruhe, Germany. [3]Osaka University, Chemistry/Biology Combined Major Program, 560-0043 Osaka, Japan.

*Present address: The Victor Chang Cardiac Research Institute – Developmental Biology Division, 2010 Sydney, Australia
‡Authors for correspondence (m.ramialison@victorchang.edu.au; jochen.wittbrodt@zoo.uni-heidelberg.de)

Accepted 22 December 2011

MATERIALS AND METHODS

Medaka stock

The iCab strain was derived by successive brother-sister inbreeding of wild-type *Oryzias latipes* from a closed stock. iCab, Cab and Heino (Loosli et al., 2000) strains were maintained as described previously (Koster et al., 1997). Embryos were staged according to Iwamatsu (Iwamatsu, 1994).

In situ hybridisation (ISH) screen

The automated ISH screen on whole-mount embryos was performed at stages 18, 24 and 32 of development as described (Quiring et al., 2004). Probes were generated from clones derived from a full-length cDNA library of 50,000 clones (sequenced by the Yunhan Hong group, National University of Singapore).

Expression data

The data present in the database were translated into a binary array of gene expression values. The substructures of a given tissue were assigned an expression value inherited from the parental tissue following the hierarchical design of the anatomical ontology. For instance, if a gene was annotated as expressed in the 'eye', all the eye substructures, i.e. 'retina', 'lens', etc., will be attributed the value 1. Genes with no expression were excluded from this array. To increase the number of expression patterns taken into account to generate the array, MEPD gene annotations for stages 18 and 19 and for stages 30 and 32 were merged, corresponding to 1 dpf and 4 dpf, respectively, as there were no significant morphological differences between these respective stages.

Identification of co-expressed genes

TERRAIN (TRN) data reduction (Saeed et al., 2003; Stuart et al., 2003; Saeed et al., 2006) was performed using the binary matrix of expression patterns as input, with the following parameters: Euclidean distance; TRN initialisation on genes; number of closest neighbours, ten. The TRN 'mountains' were extracted by individually selecting hubs of linked genes. These hubs consist of an independent group of at least four linked genes at a default distance weight threshold $\geq$0.8. The randomisation of each matrix was performed by shuffling the values within each row (gene) and by keeping the column names identical (anatomy ontology). The mean profile of each TRN mountain was calculated by averaging the expression values of the genes contained in the cluster. We then performed a hierarchical clustering of the matrix created using the EBI expression profiler (Kapushesky et al., 2004) to generate the hierarchical tree that was used to browse the co-expression network. Parameters: Euclidean distance, average linkage clustering, gene tree-based clustering.

Gene Ontology over-representation

We used ErmineJ software (Lee et al., 2005; Gillis et al., 2010) to detect shared over-represented biological terms from Gene Ontology (GO) (Ashburner et al., 2000) [*P*-value threshold t=0.05, multiple-test correction by Benjamini-Hochberg (default)]. We compiled GO annotations of medaka genes by retrieving the GO annotations of the orthologous vertebrate genes.

Sequence retrieval and orthology assignment

Medaka EST sequences from MEPD were blasted against medaka genes in Ensembl (BLASTX, e-value<10^{-7}) version 46 (Hubbard et al., 2007). For each medaka Ensembl gene, the orthology assignment by comparison with other vertebrate species was performed using the Ensembl Perl API for the Compara database, restricting the orthology search to the terms: ortholog_one2one, ortholog_one2many, inspecies_paralog and apparent_ortholog_one2one. For all the medaka genes, the non-coding regions were retrieved using the Perl API for the Ensembl Core database.

Motif search

We used the Trawler program (Ettwiller et al., 2007; Haudry et al., 2010) to identify conserved over-represented motifs in each node of the hierarchical clustering tree. Trawler was run with the default options and the number of occurrences (K) was calculated according to the number of genes composing each node (n) as: K=n/2. Orthologous sequences for the conservation calculation were retrieved from the Ensembl database.

Chromosomal clustering analysis

Perl scripts using the API Core of Ensembl v48 were written to retrieve the number of genes on each chromosome and to obtain their genomic coordinates.

Medaka synexpression group repository

We have built a repository of medaka synexpression groups, which is available at http://zooserv1.zoo.uni-heidelberg.de/mirana. Included are SGs where there are at least four genes that share the exact same occurrence of the most over-represented motifs.

Molecular cloning

Regulatory regions were cloned using the following primers (5′-3′, forward and reverse): *cdon*, TGTACGCTGCAGTTAGCGCC and AAGTGCAGGCCAGACACGCA; *otx3*, TTTGCATCCTTACACTCTGGTATTCC and ACTCTGTCCAAATAAACCAA; *hmgb2*, TGTCCTGTGTTTCTTTGCATTTGA and TGGAGCTTGACGCTTAACGG. The corresponding PCR products were subcloned into pCRII-TOPO (Invitrogen) and subsequently inserted into a pBlueScript-based transgenesis vector containing two recognition sites for the meganuclease I-*Sce*I flanking a multiple cloning site and a 3′ cassette containing a minimal promoter, enhanced GFP and an SV40 polyadenylation signal. We used the p35S minimal promoter for *cdon* and the *hsp70* minimal promoter for *hmgb2*. No minimal promoter was used for *otx3* as the regulatory region was cloned including its endogenous minimal promoter.

Transgenesis

Injections into one-cell stage medaka embryos were performed as previously described (Rembold et al., 2006). Transgenic embryos were fixed overnight at 4°C in 4% paraformaldehyde, washed three times in phosphate-buffered saline containing 0.1% Tween 20 and dechorionated. Embryos older than stage 24 were grown in 0.003% N-phenylthiourea (Sigma-Aldrich) to inhibit pigmentation. Samples were mounted using 1% SeaPlaque GTG Agarose (Lonza) and glass-bottom culture dishes (MaTek). Images were obtained using a Leica TCS SPE confocal microscope, 488 nm excitation line and ACS APO 10×/0.30 objective lens. Image stacks of up to 35 confocal slices were transformed into maximum intensity projections using ImageJ software (Version 1.41o; http://rsbweb.nih.gov/ij/). Colour balance was minimally corrected and the filter Unsharp Mask [radius, 1.0 pixels; mask weight, 0.2) was applied to all images.

RESULTS

In order to systematically identify and analyse SGs, we combined experimental and bioinformatics approaches (summarised in Fig. 1). In brief, gene expression patterns and their respective annotations were deposited in the Medaka Expression Pattern Database (MEPD). Annotation according to a standardised medaka anatomical ontology was used for clustering genes with similar expression patterns. GO annotations and non-coding sequences of medaka genes were retrieved in order to identify common biological functions and shared DNA motifs among similarly expressed genes.

Collecting spatiotemporal medaka gene expression patterns

We collected the spatiotemporal expression of randomly picked medaka genes with cDNA probes from a library of full-length sequenced clones in an automated screen on whole-mount preparations using in situ hybridisation (Fig. 1). A total of 782 clones were used to probe gene expression in embryos at three developmental time points: 1 day post-fertilisation (dpf) (late neurula stage); 2 dpf (16 somite stage); and 4 dpf (somite completion stage). Of these, the expression patterns of 407 clones were manually annotated and integrated into MEPD. By combining the data obtained from this screen and the pre-existing MEPD

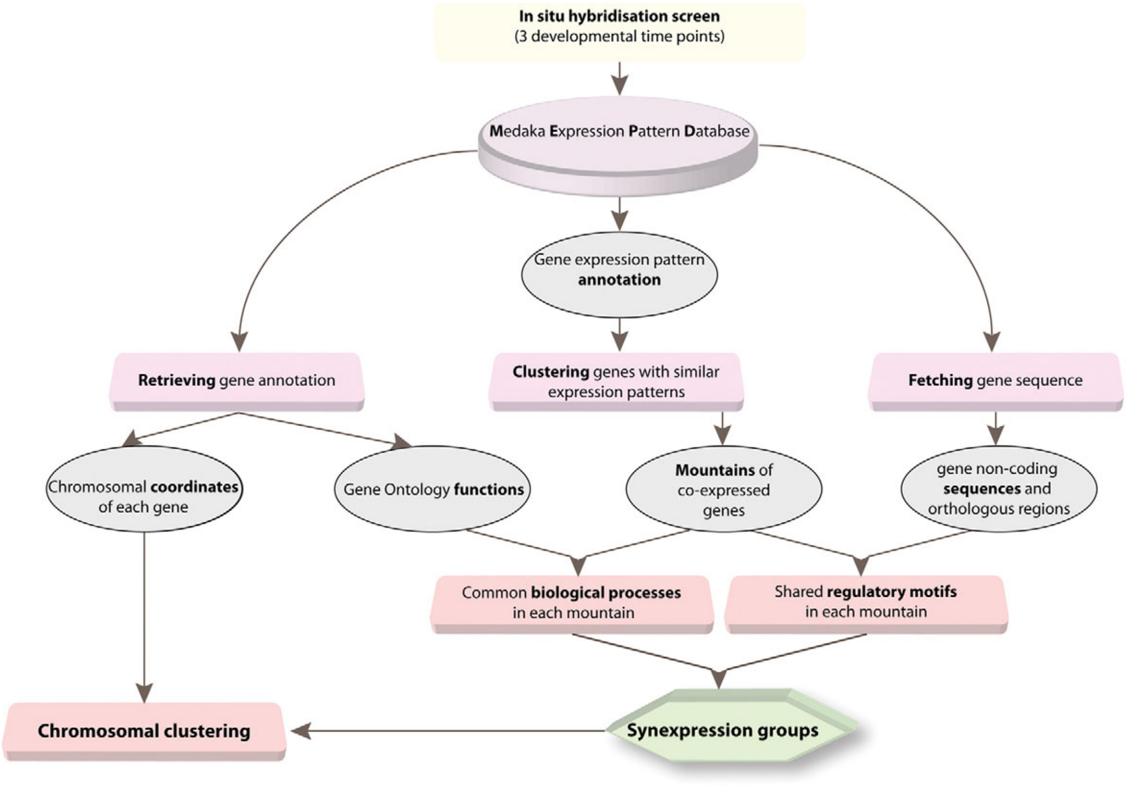
In situ hybridisation screen
(3 developmental time points)
Medaka Expression Pattern Database
Gene expression pattern annotation
Retrieving gene annotation
Clustering genes with similar expression patterns
Fetching gene sequence
Chromosomal coordinates of each gene
Gene Ontology functions
Mountains of co-expressed genes
gene non-coding sequences and orthologous regions
Common biological processes in each mountain
Shared regulatory motifs in each mountain
Chromosomal clustering
Synexpression groups

Fig. 1. Bioinformatics pipeline for the systematic discovery of medaka synexpression groups. From an automated large-scale in situ hybridisation screen on whole-mount medaka embryos at three developmental time points, images and corresponding gene expression pattern annotations were recorded in the Medaka Expression Pattern Database (MEPD). For each gene present in the database, information concerning their chromosomal locations, Gene Ontology (GO) functional annotation and upstream non-coding sequences were automatically retrieved from the Ensembl and BioMart databases. In parallel, genes sharing the same expression pattern were regrouped into 'mountains' using the TERRAIN (TRN) software. Each mountain was then assessed for common biological processes (using the ErmineJ program on GO annotations) and for common regulatory motifs (using the Trawler program on non-coding sequences). Mountains of genes sharing both biological processes and regulatory motifs were identified as synexpression groups (SGs). Finally, for each SG, gene coordinate information was used to identify collinear genes.

records from previous ISH screens (Nguyên et al., 2001; Quiring et al., 2004), the expression patterns for 750 UniGene clones, corresponding to 560 Ensembl genes, have been recorded with precise descriptions in at least one embryonic stage.

Genes ubiquitously expressed at early stages are later restricted to domains of specific co-expression

To identify groups of co-expressed genes, we performed pairwise comparisons of expression patterns documented in MEPD, using the medaka ontological annotations. Genes sharing highly similar annotations were regrouped into co-expression 'mountains' using the TERRAIN (TRN) algorithm (Saeed et al., 2003; Saeed et al., 2006) (Fig. 2); 18, 32 and 63 TRN mountains were obtained for 1, 2 and 4 dpf, respectively (supplementary material Tables S6-S8). Consistent with the tissue complexity that arises during development, we observed an increasing number of TRN mountains, reflecting emergent domains of co-expression as development proceeds. Interestingly, this increase in co-expression domains does not arise from new specific gene expression but mainly from a restriction of early ubiquitous expression into late specific expression. Indeed, we observed that although a large proportion of genes showed ubiquitous expression at early stages (58% at 1 dpf, 47 % at 2 dpf), they became spatially restricted at later stages, particularly into the developing neural tissue and central nervous system (CNS) (supplementary material Fig. S1). By 4 dpf, only 7% of the genes remain ubiquitously expressed. This specific landscape does not occur by chance, as similar analysis on randomised annotations revealed no changes in landscape (supplementary material Fig. S2; see Materials and methods).

If they shared a common domain of expression, the co-expression mountains were further grouped into 'mountain ranges' by hierarchical clustering (see Materials and methods and http://zooserv1.zoo.uni-heidelberg.de/mirana). For instance, at 4 dpf, the TRN mountain of genes expressed in 'somites' was linked to the TRN mountain of genes expressed in 'somites, pectoral fin' to form a mountain range (Fig. 2). These two TRN mountains might have a common mode of regulation as they share a common expression domain (i.e. somites) and therefore might participate in the same SG. By systematic browsing through each node of the hierarchical tree, at 4 dpf 74 nodes representing TRN mountains or mountain ranges were identified to share common expression domains representing potential SGs (supplementary material Tables S1-S4).

Shared biological processes of co-expressed genes

We investigated whether co-expressed genes belong to the same biological process by calculating GO term over-representation in the 74 groups. Of these, 30 (41%) shared statistically significant

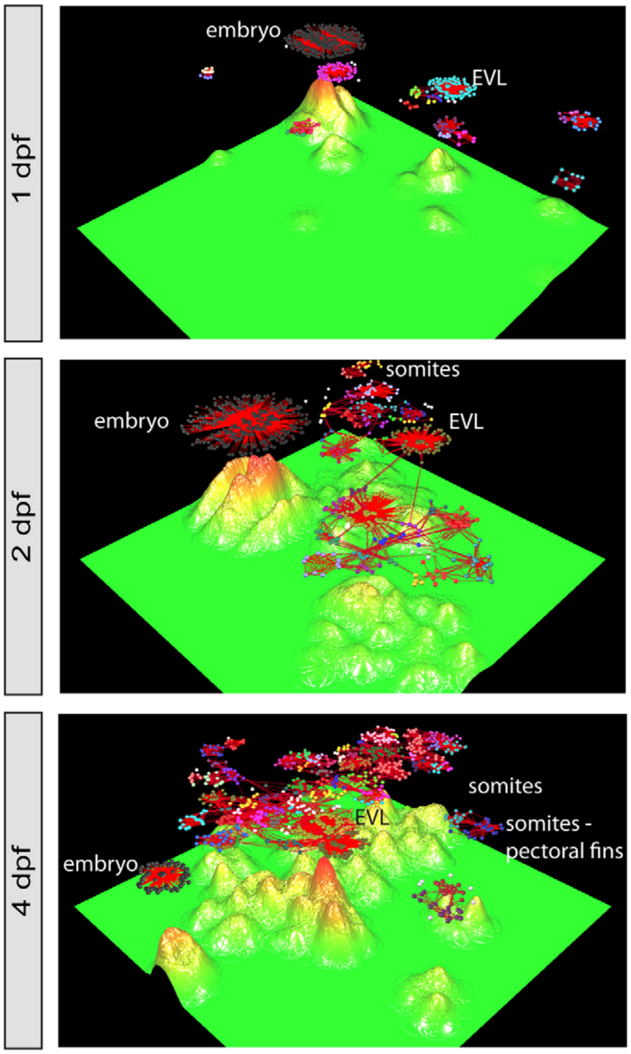

Fig. 2. Medaka co-expression TRN mountains. For each developmental stage indicated, genes (spheres) were placed in two dimensions (green field) according to the similarity of their expression pattern: the closer the spheres, the more similar the expression patterns. Beneath a cluster of spheres, a surface density (TRN mountain) was generated in a third dimension. The height of the TRN mountain represents the number of spheres it encompasses, i.e. the number of genes sharing common anatomical terms. The names of these anatomical terms are indicated: embryo, ubiquitously expressed genes; EVL, enveloping layer (see complete annotation of mountains in supplementary material Tables S6-S8). The co-expression network representing the connections between mountains was overlaid on the mountains, where nodes are genes (spheres) and edges are lines joining two spheres. Each edge represents two genes sharing one or more anatomical structures (within a mountain or across mountain ranges).

over-represented biological categories, thereby complying with the definition of SGs by Niehrs and Pollet (Niehrs and Pollet, 1999) and constituting novel medaka SGs.

We surveyed the literature to substantiate the correlation between the observed expression patterns and the over-represented GO categories (Fig. 3). The GO term 'cytoskeleton', which is over-represented in the enveloping layer (EVL) SG (Fig. 3A), reflects the active cell migration that EVL cells undergo during epiboly (Zalik et al., 1999). 'Gas exchange' (such as oxygen transport) is the primary function of the cardiovascular system (Burggren and Pinder, 1991) and is the over-represented biological process in the 'cardiovascular system' SG, which includes well-known genes such as α-*globin* (Fig. 3B). The 'contractile fibre' GO term is over-represented in the 'somites' and 'somites, pectoral fin' SGs, matching the ontogeny of these muscle cells (Fig. 3C). More strikingly, we identified an SG that contains genes that are expressed in the ciliary marginal zone (CMZ) of the retina or in the tectum proliferative zone (TPZ) of the brain (Fig. 3D). These neural domains are known to be highly proliferative and are suspected to contain stem cell niches (Wittbrodt et al., 2002; Alunni et al., 2010). Accordingly, the 'DNA-dependent DNA replication' biological process is over-represented in this SG, suggesting a role in cell proliferation as exemplified by the presence of *proliferating cell nuclear antigen* (*pcna*) in the SG.

Finally, biological categories related to ribosomal proteins ('constituent of ribosome', 'ribosomal subunit', 'cytosolic ribosome'), were the most over-represented GO categories in 18 SGs. Genes within the medaka 'ribosome biogenesis' SG were ubiquitously expressed early on and became restricted in neural derivative tissues. This dynamic spatiotemporal expression pattern corroborates the specific role of ribosomal genes during development, more particularly in pan-neural derivative tissues. Indeed, in *C. elegans* it has been suggested that ribosome biogenesis plays a role in the regulation of developmental processes (Saijou et al., 2004). In zebrafish, knockdown of 21 ribosomal proteins led to tissue-specific defects (Uechi et al., 2006). Similarly, in mouse, ribosomal protein *Rpl38* knockout also led to a tissue-specific phenotype (Kondrashov et al., 2011). Several medaka orthologues of mouse ribosomal proteins showing specific expression in that study were overlapping with the genes in our 'ribosome biogenesis' SG (namely *rpl13a*, *rpl36*, *rps4*, *rps3a* and *rps8*).

In summary, we have identified groups of co-expressed genes with a shared biological function that correlates with their expression pattern, in agreement with the definition of SGs. Some of the identified SGs have not been described previously (e.g. 'cardiovascular system'), whereas others were also found in *Xenopus*, suggesting evolutionary conservation [e.g. 'cell cycle' (Baldessari et al., 2005), 'muscle and epidermis' (Pollet et al., 2005), 'ribosome biogenesis' (Wischnewski et al., 2000)].

Synexpression groups share complex putative regulatory inputs within cis-regulatory modules

Although independent studies in specific SGs have highlighted that these genes are co-regulated by common t-AFs, the generality of these principles to other SGs remains unproven. We hypothesised that: (1) if all the genes within an SG are regulated by the same t-AF, we would expect them to harbour a common DNA sequence (or motif) in their regulatory regions that represents the binding site of the t-AF; and (2) this shared motif should be enriched in the set of genes within the same SG, as compared with its occurrence in the regulatory regions of other genes. Using the Trawler program (Ettwiller et al., 2007; Haudry et al., 2010), we systematically searched for common sequence motifs that represent likely t-AF footprints within the non-coding 5 kb upstream of the transcription start site (TSS) of genes encompassed in the 74 co-expressed TRN mountains or mountain ranges.

We identified significantly over-represented motifs de novo in 68% (50/74) of the co-expressed groups (supplementary material Tables S1, S2). About half (48%, 24/50) of these co-expressed groups fit the definition of SGs as they also share a common over-

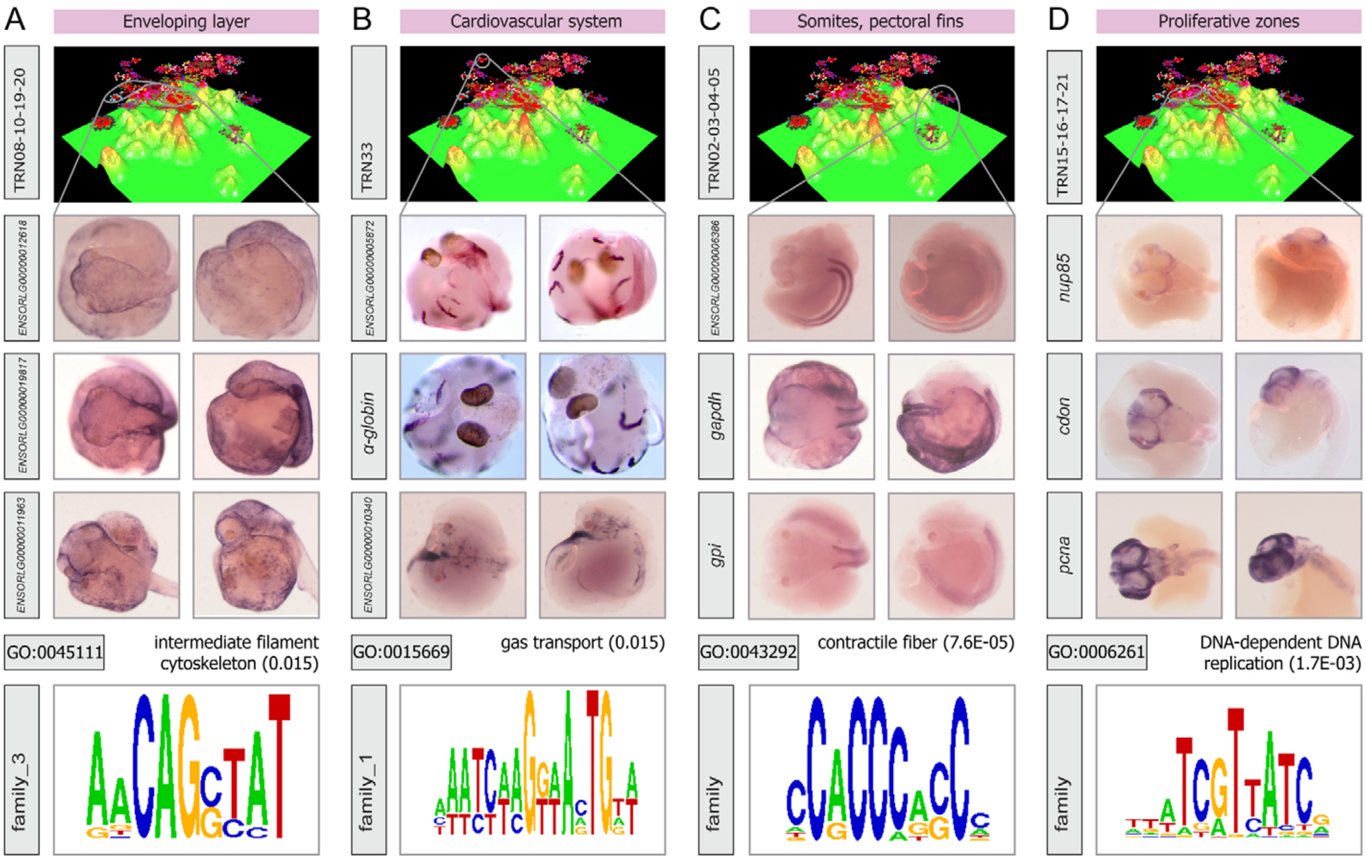

Fig. 3. Examples of synexpression groups at 4 dpf. (**A**) 'EVL', (**B**) 'cardiovascular system', (**C**) 'somites, pectoral fin' and (**D**) 'proliferative zones' SGs. (Top) Corresponding TRN mountain in the co-expression network. (Middle) Expression patterns of three representative genes from a dorsal (left) and lateral (right) view. (Bottom) Best over-represented GO term and DNA motif.

represented GO category (supplementary material Table S1). Interestingly, 52% of the co-expressed groups sharing an over-represented motif did not display an over-represented GO category (supplementary material Table S2). This latter group offers great potential for identifying new SGs because not all the medaka genes have been annotated with GO terms [only 20% (3824/18076) have been associated with a GO category to date]. We observed that SGs share a large number of over-represented motifs (20 on average), suggesting that they could be regulated by multiple t-AFs (Fig. 4A).

We examined the evolutionary conservation of these motifs by running Trawler with orthologous vertebrate sequences. Previous studies have shown that evolutionarily conserved regulatory sequences are more often present around developmental genes (Woolfe et al., 2005). As expected, we identified conserved putative regulatory motifs: 21% of shared motifs are conserved among teleosts and 1% among vertebrates (Fig. 4B). However, the largest proportion of newly discovered motifs were medaka specific (78%), which suggests that non-conserved DNA motifs can potentially have a regulatory function.

To further investigate the functionality of this shared input, we focused on the proliferative zone SG for experimental validation. We noticed that the positions of the shared motifs along the upstream regions of the genes are not random but are rather clustered into modules (Fig. 5, light-pink squares). To confirm this observation, we investigated whether there is a bias in the distances between the positions of two consecutive motifs, or whether they are randomly distributed along the putative regulatory region. Comparing the distances between motifs in the SG against distances in a dataset of random positions, we observed an enrichment of clustered motifs separated by short distances (less than 100 bp) in the SG (Fig. 4C). This modular organisation of the over-represented motifs in SGs supports their regulatory potential, as transcription factors are well known to interact with other co-factors, likely aided by colocalisation into compact CRMs (Howard and Davidson, 2004; Wasserman and Sandelin, 2004). Moreover, the multiple occurrence of a de novo identified motif within the same CRM is in agreement with previous studies that have shown that local clustering of transcription factor binding sites is a feature of vertebrate enhancers (Gotea et al., 2010).

To gain further insight into the relevance of the newly identified motifs, we investigated the specific signalling pathways involved in the regulation of the 'proliferative' SG and attempted to link these pathways with specific instances of the de novo motifs. On the one hand, we performed a signalling pathway enrichment analysis using DAVID (version 6.7) (Huang et al., 2009). The first enriched pathway from the KEGG database was 'DNA replication', confirming our previous GO analysis (supplementary material Fig. S3A). Interestingly, the first enriched pathway from the BIOCARTA database is the 'p53 signalling pathway', of which three well-described genes, namely *e2f*, *cdk2* and *pcna*, are present in this SG (supplementary material Fig. S3B). On the other hand, the first over-represented motif identified with Trawler (Fig. 3D, lower panel; supplementary material Fig. S3C, upper panel) was predicted to

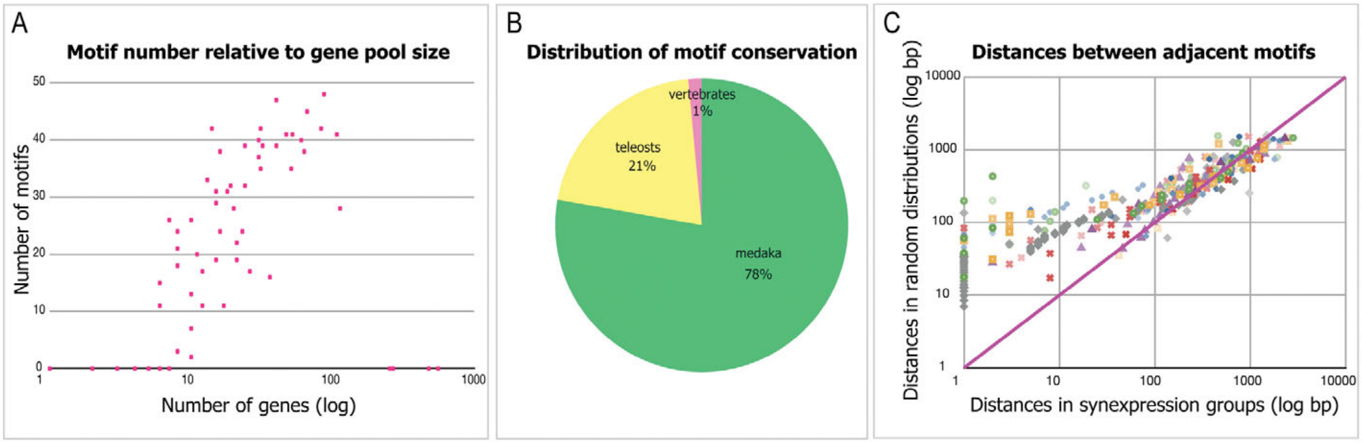

Fig. 4. Properties of shared DNA motifs within a synexpression group. (**A**) Distribution of the number of motifs discovered for each SG according the number of regulatory regions investigated. (**B**) Distribution of conservation scores of the over-represented motifs. (**C**) Quantile-quantile plot representing the distribution of distances between two consecutive over-represented motifs in each of the SGs (coloured geometrical shapes) compared with a dataset of randomly generated positions containing the same number of positions as in the SG (obtained after 100 randomisations).

overlap with the known Foxl1 binding site (MA0033 from the Jaspar database) (Portales-Casamar et al., 2010) (supplementary material Fig. S3C, lower panel), a transcription factor belonging to the class O forkhead box family. Strikingly, several studies have suggested a parallel role of the FOXO and p53 pathways: crosstalk between the p53 and FOXO transcription factors has been described and they have been demonstrated to share several common target genes (You and Mak, 2005). Taken together, our motif discovery and pathway analyses suggest that Foxl1 is a likely candidate to regulate genes belonging to the p53 signalling pathway. This example highlights the value of our pipeline to generate robust hypotheses for further experimental validation.

The regulatory activity of predicted cis-regulatory modules recapitulates subdomains of endogenous gene expression

To experimentally validate that the CRMs encompassing shared DNA motifs can drive specific expression in related tissues, we assayed the activity of six CRMs from genes within the ‘proliferation’ SG (*cdon*, *e2f*, *hmgb2*, *nup85*, *otx3* and *rpf20*). Three of them drove expression in the proliferative zones (Fig. 6).

Cdon (or *Cdo*, cell adhesion molecule-related/down-regulated by oncogenes) has been shown to promote the differentiation of several cell lineages in mouse (Oh et al., 2009), where it is initially expressed in the CNS and sensory organs and becomes restricted to the dorsal ventricular part of the brain (Mulieri et al., 2000), which is known to contain proliferative cells. This expression pattern is similar to that of medaka *cdon*, which is specifically restricted to the proliferative zones of the retina and brain at later stages of development. *cdon* harbours one conserved predicted motif in a region proximal to the TSS (Fig. 6A, red cross). We tested the regulatory activity of a 511 bp conserved region (Fig. 6A, green rectangle) containing this motif. We found that it recapitulates the endogenous expression pattern in the forebrain proliferative zone (FPZ) and midbrain-hindbrain boundary at 2 dpf, and also in the posterior part of the FPZ at 4 dpf (Fig. 6A). This exemplifies the ability of an evolutionarily conserved CRM encompassing the newly identified over-represented motif to drive specific spatiotemporal expression.

otx3 (or *dmbx1*) is a transcriptional repressor that plays a crucial role in brain morphogenesis (Zhang et al., 2002; Kimura et al., 2005). In zebrafish, it has been shown to regulate cell cycle exit during neurogenesis (Wong et al., 2010). This function correlates with its specific expression in the TPZ of the brain, which represents an area of proliferating as well as post-mitotic cells (Nguyên et al., 2001). Two conserved regions were identified in the upstream regulatory region of *otx3*, but only the region distal to the TSS contained an over-represented motif as predicted by Trawler (Fig. 6B, green circle). When we tested the regulatory activity of a 3 kb fragment (Fig. 6B, green rectangle) that contains both conserved regions, it recapitulated the endogenous expression pattern of *otx3* in the TPZ. Deletion of the region with the over-represented motif, however, results in the complete loss of reporter expression (data not shown). In this case, a stretch of conserved DNA does not necessarily drive expression, but the conserved fragment that contains the identified over-represented motif is necessary to successfully recapitulate the endogenous expression pattern.

hmgb2 (*high mobility group 2*) is a cell proliferation promoting factor implicated in several oncogenic processes (Kwon et al., 2010). In accordance with this function, it is strongly and specifically expressed in all proliferative areas of the medaka embryo: the CMZ of the retina, the FPZ and TPZ of the brain, the rhombic lips and the marginal zones of the fins (Fig. 6C). In this case, we tested the activity of a 696 bp non-conserved CRM containing two copies of the same motif (Fig. 6C, purple triangles) as predicted by Trawler. This short fragment recapitulates a subset of the endogenous expression pattern in the anterior part of the FPZ (Fig. 6C) and in the ventral CMZ, the otic vesicles and pectoral fins (data not shown).

For the three other CRMs tested, the *e2f* CRM recapitulated only transiently the early ubiquitous expression pattern, the *nup85* CRM did not drive a specific expression pattern and the *rpf2* CRM drove an expression pattern different to that of the *rpf2* gene itself. It is possible that these CRMs regulate other genes or act in synergy with other CRMs to perform their regulatory function.

The three positive examples shown above indicate that half of the computationally predicted modules containing de novo over-represented motifs drive specific gene expression.

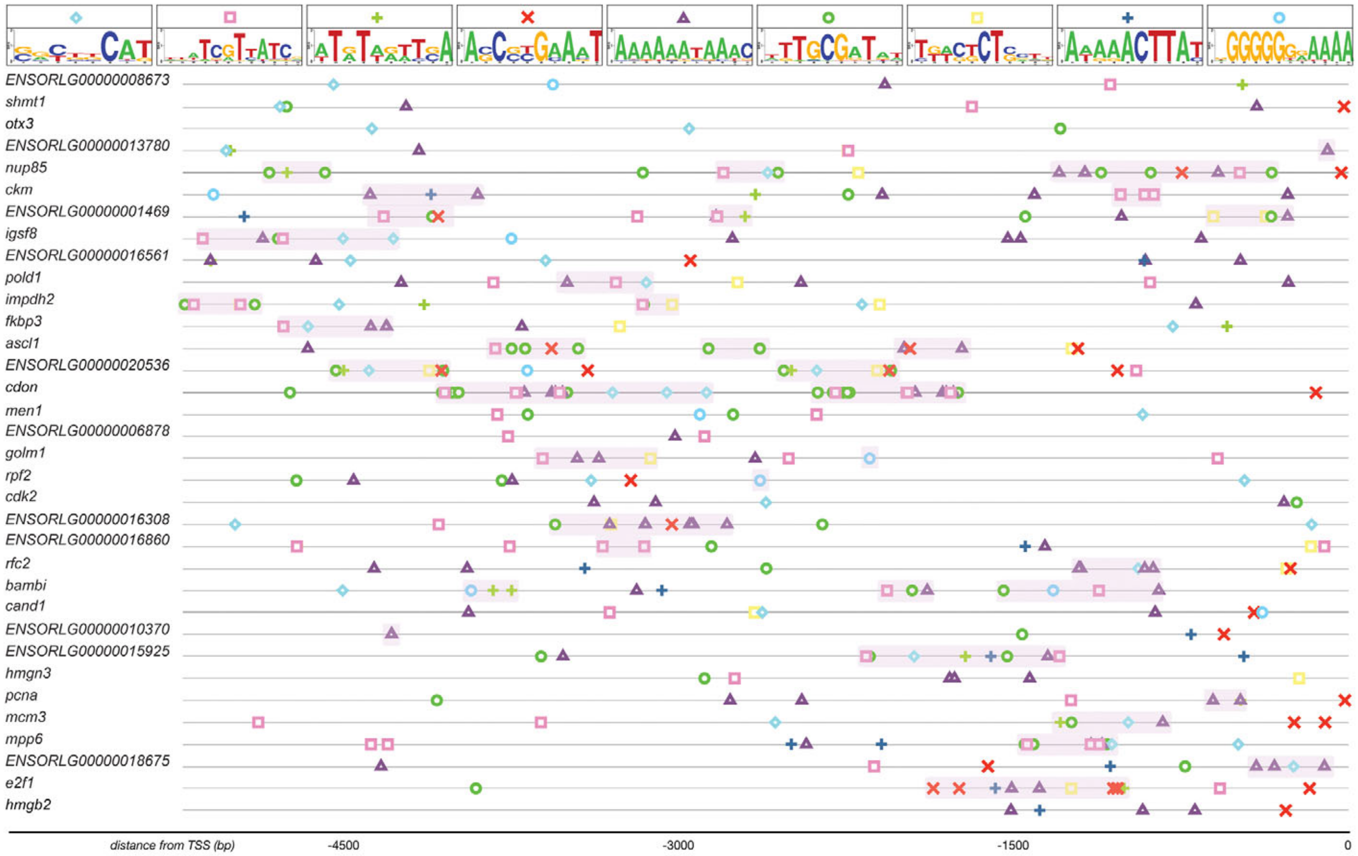

Fig. 5. De novo discovery of shared DNA motifs in the 'proliferative zones' synexpression group. DNA motifs discovered by Trawler are displayed along the 5 kb region upstream of the transcription start site (TSS) of the genes in this SG. Sixteen overlapping motifs originally identified by Trawler were subsequently merged using STAMP software (Mahony and Benos, 2007) into nine distinct motifs as represented by a coloured geometrical shape (top). Cis-regulatory modules (CRMs) are highlighted in the pink-shaded areas, consisting of clustered DNA motifs separated by less than 300 bp.

Synexpression groups contain instances of genes that are clustered at the same chromosomal location

It has been generally assumed that genes belonging to the same SG can be positioned anywhere in the genome, provided that they share a common regulatory input. We sought to systematically investigate whether this is a general property of SGs by examining the genomic location of genes within each of the 50 SGs that share over-represented DNA motifs. To our surprise, 30% (15/50) of 50 SGs sharing a common regulatory input contained pairs of genes located on the same chromosome and separated by fewer than ten genes (supplementary material Table S5). This spatial segregation is statistically significant as compared with random pairs of genes on the same chromosome (*P*<0.03 for 100,000 iterations, except for one pair of genes located on ultracontig88, where only 21 genes were annotated).

To further address the significance of this arrangement, we investigated whether the other genes lying between or near co-expressed pairs of genes also exhibit a similar expression pattern (Fig. 7; supplementary material Fig. S4). Six chromosomal locations (supplementary material Table S5, bold) revealed strikingly similar expression patterns in the newly investigated proximal genes. In particular, two distinct locations with clusters of genes expressed in the proliferative zones (Fig. 7A; supplementary material Fig. S4A) revealed four additional genes displaying identical expression patterns. In one of these clusters, the expression pattern of one gene (*ENSORLG0000001937*) is inverted, i.e. it is specifically repressed in the proliferative zones (supplementary material Fig. S4A), hinting at a common regulation pathway with the potential for both positive and negative effects on gene expression. Two loci include two gene pairs expressed in the 'somites, pectoral fin' SG (Fig. 7B; supplementary material Fig. S4B). One contains paralogous Troponin genes but also includes *myoD*, the role of which in muscle formation is very well characterised (Rudnicki et al., 1993). The expression of *myoD* in the somites was confirmed in the second round of ISH along with that of *rfwd3* (*ring finger and WD repeat domain 3*), which is part of this cluster but has never been shown to be involved in muscle formation (Fig. 7B). Two additional genes were further confirmed by ISH to be co-expressed in the somites in the second locus (supplementary material Fig. S4B). ISH also highlighted co-expression of two intervening genes in a cluster of four genes previously grouped into the 'EVL' SG on chromosome 1 (Fig. 7C). Finally, two genes belonging to the 'retina, forebrain, midbrain, cerebellum, pectoral fin' SG (supplementary material Table S5) and located together on chromosome 12 (Fig. 7D) display striking similarity of expression, especially in the otic vesicle structure. These two genes are separated by a single gene, *gfm2*, which is expressed in the otic vesicle domain as judged by ISH analysis (Fig. 7D, white arrows).

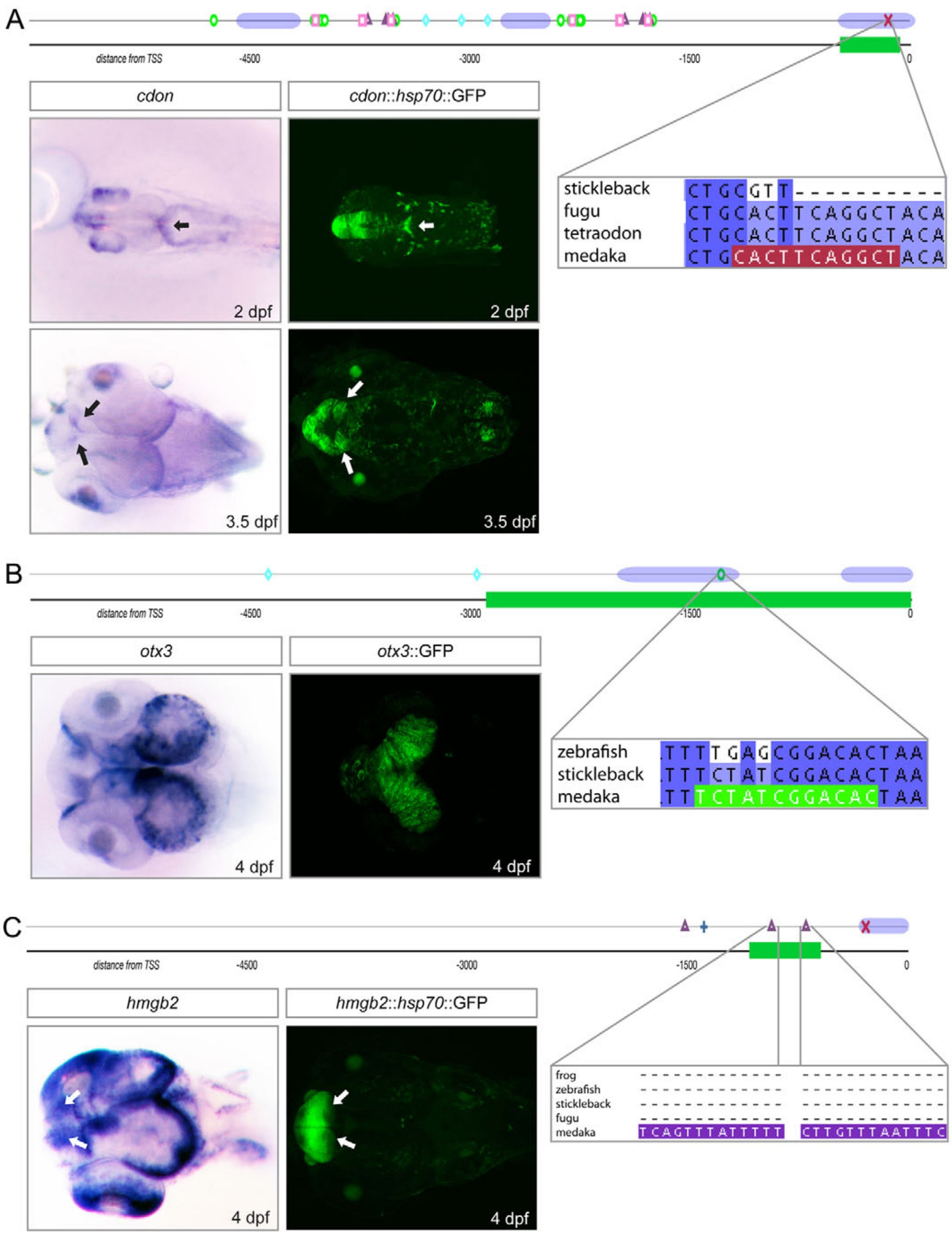

Fig. 6. Transgenic reporter assays in the 'proliferative zones' synexpression group. (Top) The 5 kb sequence upstream of the TSS of three examples of genes showing over-represented motifs (coloured geometrical shapes), evolutionarily conserved regions (blue oval) and regions selected for transgenic assays (green rectangles). (Right) The conservation pattern of the de novo discovered motif in the region tested. (Left) Endogenous expression of the gene examined by ISH (dorsal view). (Middle) GFP expression pattern in transgenic assays (dorsal view). Reporter construct and endogenous gene show co-expression (arrows) in (**A**) the anterior part of the forebrain and in the midbrain-hindbrain boundary at 2 dpf and in the posterior part of the forebrain proliferative zone (FPZ) at 4 dpf for *cdon*, (**B**) the tectum marginal zone of the brain for *otx3*, (**C**) and in the anterior part of the FPZ for *hmgb2*.

One explanation for this chromosomal clustering is that these genes arose by gene duplication (Brunet et al., 2006). We therefore investigated whether the clustered genes in SGs share a common evolutionary origin and found that 12 SGs contain clustered co-expressed genes without any structural or ancestral relation. Only four SGs contain paralogous genes (supplementary material Table S5, italics); namely, the 'yolk syncytial layer' (YSL) SG with tandemly duplicated Apolipoprotein genes, the 'cardiovascular system' SG with a Globin gene cluster that arose from duplication of ancestral genes (Hardison and Miller, 1993), the 'somites' SG with a group of Troponin genes that are also arranged into clusters of paralogous pairs in human (Cullen et al., 2004) (Fig. 7B), and the 'EVL' SG cluster, with four of six uncharacterised genes sharing a conserved protein domain suggesting a paralogous relationship (Fig. 7C).

As the majority of synexpression clusters did not arise by gene duplication, we examined whether this spatial arrangement was selected under an evolutionary constraint by investigating the synteny conservation within the clusters. Four SG clusters display noteworthy synteny conservation between medaka and mouse [two 'proliferative' SGs (Fig. 7A; supplementary material Fig. S4A) and the 'retina, forebrain, midbrain, cerebellum, pectoral fin' SG (Fig. 7D)]. This conservation extends to human in the 'somites, pectoral fin' SG (supplementary material Fig. S4B). This locus is syntenic with human chromosome 19, which is known to harbour a concentration of skeletal muscle-specific transcripts (Bortoluzzi et al., 1998).

To reinforce the relevance of this genomic clustering, we investigated whether the expression pattern of the genes conserved in synteny is likewise conserved by screening the zebrafish and

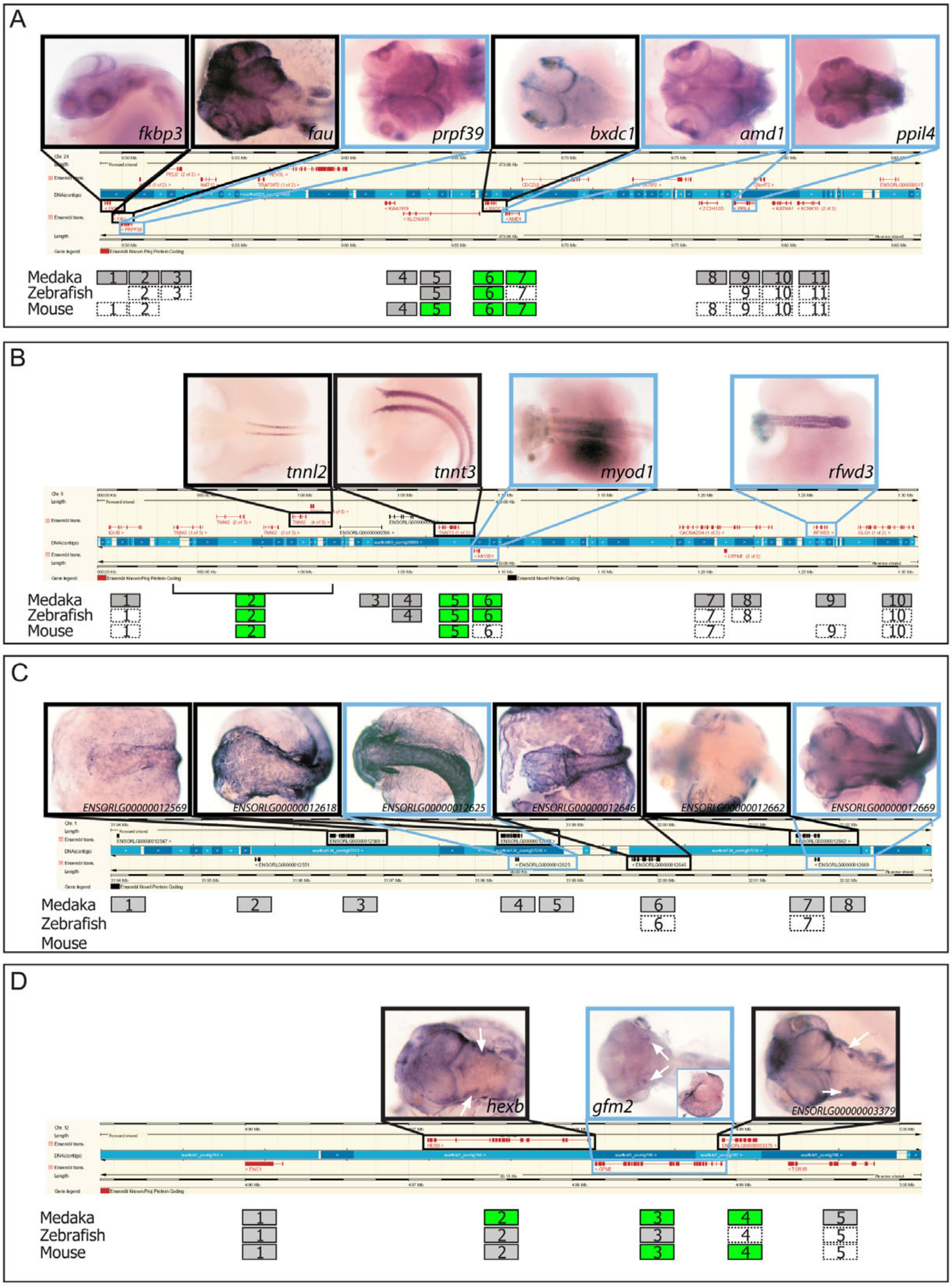

Fig. 7. Chromosomal clusters of synexpression groups and synteny conservation. Snapshots of medaka chromosomal loci harbouring co-expressed genes from the Ensembl genome browser. Dorsal views of initial pairs of co-expressed genes already present in MEPD (black squares) and newly investigated gene expression patterns (blue squares). Genes are numbered according to their position along the chromosome from 5′ to 3′ (regardless of the strand, except for A which is reverse strand only). Grey rectangles represent genes with conserved synteny. Green rectangles represent genes with conserved synteny and conserved expression patterns across the different species represented. Dotted rectangles represent genes on different chromosomes in the other species. (**A**) Synexpression cluster of genes co-expressed in the ciliary marginal zone of the retina and the tectum proliferative zone. (**B**) Synexpression clusters of genes co-expressed in the somites and pectoral fins. (**C**) Synexpression clusters of genes co-expressed in the EVL. (**D**) Synexpression clusters of genes co-expressed in retina, forebrain, midbrain and cerebellum with strong specific expression in the otic vesicles (arrows).

mouse expression pattern repositories [ZFIN (Bradford et al., 2011), GXD (Finger et al., 2011), Eurexpress (Diez-Roux et al., 2011)]. Strikingly, in all of the three clusters that show conserved synteny, we found at least two genes that display similar expression patterns in mouse (Fig. 7A,B,D, genes in green rectangles). For instance, genes expressed in the TPZ in medaka and zebrafish are also expressed in the subventricular zone in mouse, a region that is known to contain neural stem cells. Only two synexpression clusters display poor synteny conservation: the 'somite' SG contains only two genes with conserved synteny compared with mouse, which are the known Troponin paralogues (Fig. 7B), and within the 'EVL' SG the synteny conservation could not be assessed as no orthologous genes were found, suggesting that this is a cluster of medaka-specific genes (Fig. 7C).

In summary, we have shown that 30% of SGs include genes that are clustered at the same chromosomal locus, possibly representing an evolutionarily conserved feature along with the conservation of gene expression.

DISCUSSION

Towards a comprehensive collection of vertebrate developmental gene expression patterns

Although this represents a pilot study, it serves as a pioneer work to investigate general rules of SG regulation. The whole-genome study of SGs is mainly constrained by technical limitations in the simultaneous recording of precise spatiotemporal expression patterns during development. DNA microarray assays, for instance, have been successfully applied for studying genome-wide transcriptional regulation (Furlong et al., 2001; Baldessari et al., 2005) as they allow quantitative and temporal monitoring of expression during embryonic development, but lack the three-dimensional information concerning tissue connectivity within the embryo. Large-scale ISH on whole-mount embryos is an approach that can provide spatial information at high resolution for gene expression on an entire organism level (Lynch et al., 1995; Gawantka et al., 1998; Kawashima et al., 2000; Neidhardt et al., 2000; Kudoh et al., 2001; Pollet et al., 2005; Lecuyer et al., 2007). In *Xenopus*, microarray analysis not only confirmed SGs identified by ISH, but also allowed the discovery of novel SGs. Nevertheless, because they provide a low-resolution analysis of gene expression patterns compared with ISH, the risk of including false positives will increase (Baldessari et al., 2005). Studies combining both techniques would be ideal, as demonstrated in invertebrate models (Tomancak et al., 2002). Medaka is well suited as a vertebrate model organism on which to perform such an analysis because of its amenable properties for high-throughput screens (such as rapid and external development, embryo transparency) (Wittbrodt et al., 2002) combined with the completed assembly of its compact genome (Kasahara et al., 2007). The development of virtual embryos (Bryson-Richardson et al., 2007; Keller et al., 2008; Richardson et al., 2010) and subsequent three-dimensional expression pattern recordings offers great promise to accelerate the analysis of SGs.

A multi-combinatorial regulatory input that leads to co-expression

We discovered several different DNA motifs that are shared between genes of the same SGs, although the combinatorial input differs from one gene to another. This observation is in accordance with several studies performed in other developmental model organisms that propose that similar expression patterns are directed by different regulatory codes. In *Ciona*, precise dissection of the regulation of muscle-specific genes highlighted a great variability in the number and types of shared regulatory motifs that are able to drive co-expression in this tissue (Brown et al., 2007). In *D. melanogaster*, combinations of binding events that direct gene co-expression were found to be variable in occupancy and intensity (Zinzen et al., 2009).

Nonetheless, this diversity of regulatory input might confer robustness against perturbations owing to the redundancy of the combinatorial input. Small changes in combinations of regulatory input or mutations in the regulatory repertoire might not be lethal for the embryo, but could trigger subtle expression pattern changes that would translate into phenotypic changes leading to animal diversity.

Collinearity in synexpression groups?

A key finding in our study is the identification of 12 genomic loci at which structurally non-related genes belonging to the same SG were clustered. Since this work only represents a pilot study, the prospect for identifying further synexpression clusters is high. The analysis of tissue-specific EST collections showed concentrations of multiple genes at the same loci (Bortoluzzi et al., 1998; Ko et al., 1998; Li et al., 2005; Vogel et al., 2005). Genes with common biological properties, such as highly expressed genes (Caron et al., 2001), housekeeping genes (Lercher et al., 2002), genes involved in the same biological pathway (Lee and Sonnhammer, 2003) and cancer genes (Zhou et al., 2003), were also shown to be clustered at the same chromosomal loci. The mechanisms that promote this chromosomal clustering remain largely unknown. Bi-directional promoters could explain why pairs of genes are co-expressed but do not explain why more than two genes share similar expression patterns (Cho et al., 1998; Cohen et al., 2000).

SGs have been considered as the eukaryotic counterparts of bacterial operons (Niehrs and Pollet, 1999). By analogy, co-expressed genes might be maintained along the same chromosome, owing to a single shared regulatory input, in order to lower the amount of regulatory information (Price et al., 2005). On the contrary, our data suggest that the regulatory information is redundant and combinatorial in SGs with genes clustered at the same locus. It is likely that the two mechanisms of regulation are not mutually exclusive, and that synexpression members are clustered at the same locus owing to a common mode of regulation. It is possible that they are co-regulated by global control regions (Spitz et al., 2003) that reside in a distal enhancer. Owing to computational restrictions, we limited our analysis to look for common over-represented motifs in the 5 kb upstream regions of the SG genes. In the future, increasing computational power will overcome this restriction. Given that the synexpression clusters display striking synteny and similar expression patterns, it is tempting to speculate that they might also show a sequential onset of gene expression, similar to the collinearity described for Hox gene clusters (Duboule, 1998).

Conclusion

We have performed the first systematic investigation of the regulatory properties of SGs. We initially identified medaka SGs and thereafter revealed that SGs share a common complex input that is potentially regulatory and embedded within CRMs. We also provide the first evidence that a significant proportion of SGs contain genes that are closely clustered at genomic loci. De novo unbiased approaches such as ours will aid in deciphering the complexity of interactions that occur in the regulation of gene co-expression during development.

Acknowledgements
We thank the J.W. laboratory for their valuable input on this project; Stuart Archer, Alvaro Mateos-Gil and Steffen Lemke for critical comments on the manuscript; Yunhan Hong for end-sequencing the 50 K medaka library; and Franck Bourrat for his valuable contribution to the annotations.

Funding
This project was supported by the European Union Sixth Framework Programme (FP6) STREP Plurigenes (J.W.) and Deutsche Forschungsgemeinschaft (DFG) SFB 488 (J.W.).

Competing interests statement
The authors declare no competing financial interests.

Author contributions
M.R., T.H. and J.W. designed experiments; M.R., R.R., T.H., B.W., T.K. and C.M.L. conducted experiments; M.R. analysed the data and wrote the paper with contributions from R.R., T.H. and J.W.

Supplementary material
Supplementary material available online at
http://dev.biologists.org/lookup/suppl/doi:10.1242/dev.071803/-/DC1

References

Alunni, A., Hermel, J. M., Heuze, A., Bourrat, F., Jamen, F. and Joly, J. S. (2010). Evidence for neural stem cells in the medaka optic tectum proliferation zones. *Dev. Neurobiol.* **70**, 693-713.

Amaya, E. (2005). Xenomics. *Genome Res.* **15**, 1683-1691.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29.

Baldessari, D., Shin, Y., Krebs, O., König, R., Koide, T., Vinayagam, A., Fenger, U., Mochii, M., Terasaka, C., Kitayama, A. et al. (2005). Global gene expression profiling and cluster analysis in Xenopus laevis. *Mech. Dev.* **122**, 441-475.

Bortoluzzi, S., Rampoldi, L., Simionati, B., Zimbello, R., Barbon, A., d'Alessi, F., Tiso, N., Pallavicini, A., Toppo, S., Cannata, N. et al. (1998). A comprehensive, high-resolution genomic transcript map of human skeletal muscle. *Genome Res.* **8**, 817-825.

Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D. G., Knight, J., Mani, P., Martin, R., Moxon, S. A. et al. (2011). ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.* **39**, D822-D829.

Brown, C. D., Johnson, D. S. and Sidow, A. (2007). Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317**, 1557-1560.

Brunet, F. G., Roest Crollius, H., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., Laudet, V. and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* **23**, 1808-1816.

Bryson-Richardson, R. J., Berger, S., Schilling, T. F., Hall, T. E., Cole, N. J., Gibson, A. J., Sharpe, J. and Currie, P. D. (2007). FishNet: an online database of zebrafish anatomy. *BMC Biol.* **5**, 34.

Burggren, W. W. and Pinder, A. W. (1991). Ontogeny of cardiovascular and respiratory physiology in lower vertebrates. *Annu. Rev. Physiol.* **53**, 107-135.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A. et al. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289-1292.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65-73.

Cohen, B. A., Mitra, R. D., Hughes, J. D. and Church, G. M. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **26**, 183-186.

Cullen, M. E., Dellow, K. A. and Barton, P. J. (2004). Structure and regulation of human troponin genes. *Mol. Cell. Biochem.* **263**, 81-90.

Diez-Roux, G., Banfi, S., Sultan, M., Geffers, L., Anand, S., Rozado, D., Magen, A., Canidio, E., Pagani, M., Peluso, I. et al. (2011). A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.* **9**, e1000582.

Duboule, D. (1998). Vertebrate hox gene regulation: clustering and/or colinearity? *Curr. Opin. Genet. Dev.* **8**, 514-518.

Ernsberger, U. (2000). Evidence for an evolutionary conserved role of bone morphogenetic protein growth factors and phox2 transcription factors during noradrenergic differentiation of sympathetic neurons. Induction of a putative synexpression group of neurotransmitter-synthesizing enzymes. *Eur. J. Biochem.* **267**, 6976-6981.

Ettwiller, L., Paten, B., Ramialison, M., Birney, E. and Wittbrodt, J. (2007). Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods* **4**, 563-565.

Finger, J. H., Smith, C. M., Hayamizu, T. F., McCright, I. J., Eppig, J. T., Kadin, J. A., Richardson, J. E. and Ringwald, M. (2011). The mouse Gene Expression Database (GXD): 2011 update. *Nucleic Acids Res.* **39**, D835-D841.

Furlong, E. E., Andersen, E. C., Null, B., White, K. P. and Scott, M. P. (2001). Patterns of gene expression during Drosophila mesoderm development. *Science* **293**, 1629-1633.

Gawantka, V., Pollet, N., Delius, H., Vingron, M., Pfister, R., Nitsch, R., Blumenstock, C. and Niehrs, C. (1998). Gene expression screening in Xenopus identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. *Mech. Dev.* **77**, 95-141.

Gillis, J., Mistry, M. and Pavlidis, P. (2010). Gene function analysis in complex data sets using ErmineJ. *Nat. Protoc.* **5**, 1148-1159.

Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A. and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* **20**, 565-577.

Grade, C. V. C., Salerno, M. S., Schubert, F. R., Dietrich, S. and Alvares, L. E. (2009). An evolutionarily conserved Myostatin proximal promoter/enhancer confers basal levels of transcription and spatial specificity in vivo. *Dev. Genes Evol.* **219**, 497-508.

Hardison, R. and Miller, W. (1993). Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Mol. Biol. Evol.* **10**, 73-102.

Haudry, Y., Ramialison, M., Paten, B., Wittbrodt, J. and Ettwiller, L. (2010). Using Trawler_standalone to discover overrepresented motifs in DNA and RNA sequences derived from various experiments including chromatin immunoprecipitation. *Nat. Protoc.* **5**, 323-334.

Henrich, T., Ramialison, M., Quiring, R., Wittbrodt, B., Furutani-Seiki, M., Wittbrodt, J., Kondoh, H. and Database, M. E. P. (2003). MEPD: a Medaka gene expression pattern database. *Nucleic Acids Res.* **31**, 72-74.

Henrich, T., Ramialison, M., Wittbrodt, B., Assouline, B., Bourrat, F., Berger, A., Himmelbauer, H., Sasaki, T., Shimizu, N., Westerfield, M. et al. (2005). MEPD: a resource for medaka gene expression patterns. *Bioinformatics* **21**, 3195-3197.

Howard, M. L. and Davidson, E. H. (2004). cis-Regulatory control circuits in development. *Dev. Biol.* **271**, 109-118.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44-57.

Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. et al. (2007). Ensembl 2007. *Nucleic Acids Res.* **35**, D610-D617.

Iwamatsu, T. (1994). Stages of normal development in the medaka Oryzias latipes. *Zool. Sci.* **11**, 825-839.

Kapushesky, M., Kemmeren, P., Culhane, A. C., Durinck, S., Ihmels, J., Korner, C., Kull, M., Torrente, A., Sarkans, U., Vilo, J. et al. (2004). Expression Profiler: next generation-an online platform for analysis of microarray data. *Nucleic Acids Res.* **32**, W465-W470.

Karaulanov, E., Knöchel, W. and Niehrs, C. (2004). Transcriptional regulation of BMP4 synexpression in transgenic Xenopus. *EMBO J.* **23**, 844-856.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y. et al. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719.

Kawashima, T., Kawashima, S., Kanehisa, M., Nishida, H. and Makabe, K. W. (2000). MAGEST: MAboya gene expression patterns and sequence tags. *Nucleic Acids Res.* **28**, 133-135.

Keller, P. J., Schmidt, A. D., Wittbrodt, J. and Stelzer, E. H. (2008). Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science* **322**, 1065-1069.

Kimura, K., Miki, T., Shibasaki, T., Zhang, Y., Ogawa, M., Saisho, H., Okuno, M., Iwanaga, T. and Seino, S. (2005). Functional analysis of transcriptional repressor Otx3/Dmbx1. *FEBS Lett.* **579**, 2926-2932.

Ko, M. S., Threat, T. A., Wang, X., Horton, J. H., Cui, Y., Pryor, E., Paris, J., Wells-Smith, J., Kitchen, J. R., Rowe, L. B. et al. (1998). Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the X chromosome. *Hum. Mol. Genet.* **7**, 1967-1978.

Kondrashov, N., Pusic, A., Stumpf, C. R., Shimizu, K., Hsieh, A. C., Xue, S., Ishijima, J., Shiroishi, T. and Barna, M. (2011). Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. *Cell* **145**, 383-397.

Koster, R., Stick, R., Loosli, F. and Wittbrodt, J. (1997). Medaka spalt acts as a target gene of hedgehog signaling. *Development* **124**, 3147-3156.

Kudoh, T., Tsang, M., Hukriede, N. A., Chen, X., Dedekian, M., Clarke, C. J., Kiang, A., Schultz, S., Epstein, J. A., Toyama, R. et al. (2001). A gene expression screen in zebrafish embryogenesis. *Genome Res.* **11**, 1979-1987.

Kwon, J. H., Kim, J., Park, J. Y., Hong, S. M., Park, C. W., Hong, S. J., Park, S. Y., Choi, Y. J., Do, I., Joh, J. W. et al. (2010). Overexpression of HMGB2 is associated with tumor aggressiveness and prognosis of hepatocellular carcinoma. *Clin. Cancer Res.* **16**, 5511-5521.

Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P. and Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **131**, 174-187.

Lee, H. K., Braynen, W., Keshav, K. and Pavlidis, P. (2005). ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* **6**, 269.

Lee, J. M. and Sonnhammer, E. L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* **13**, 875-882.

Lercher, M. J., Urrutia, A. O. and Hurst, L. D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31**, 180-183.

Li, Q., Lee, B. T. and Zhang, L. (2005). Genome-scale analysis of positional clustering of mouse testis-specific genes. *BMC Genomics* **6**, 7.

Loosli, F., Koster, R. W., Carl, M., Kuhnlein, R., Henrich, T., Mucke, M., Krone, A. and Wittbrodt, J. (2000). A genetic screen for mutations affecting embryonic development in medaka fish (Oryzias latipes). *Mech. Dev.* **97**, 133-139.

Lynch, A. S., Briggs, D. and Hope, I. A. (1995). Developmental expression pattern screen for genes predicted in the C. elegans genome sequencing project. *Nat. Genet.* **11**, 309-313.

Mahony, S. and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* **35**, W253-W258.

Mulieri, P. J., Okada, A., Sassoon, D. A., McConnell, S. K. and Krauss, R. S. (2000). Developmental expression pattern of the cdo gene. *Dev. Dyn.* **219**, 40-49.

Neidhardt, L., Gasca, S., Wertz, K., Obermayr, F., Worpenberg, S., Lehrach, H. and Herrmann, B. G. (2000). Large-scale screen for genes controlling mammalian embryogenesis, using high-throughput gene expression analysis in mouse embryos. *Mech. Dev.* **98**, 77-94.

Nguyên, V., Joly, J. and Bourrat, F. (2001). An in situ screen for genes controlling cell proliferation in the optic tectum of the medaka (Oryzias latipes). *Mech. Dev.* **107**, 55-67.

Niehrs, C. and Pollet, N. (1999). Synexpression groups in eukaryotes. *Nature* **402**, 483-487.

Oh, J.-E., Bae, G.-U., Yang, Y.-J., Yi, M.-J., Lee, H.-J., Kim, B.-G., Krauss, R. S. and Kang, J.-S. (2009). Cdo promotes neuronal differentiation via activation of the p38 mitogen-activated protein kinase pathway. *FASEB J.* **23**, 2088-2099.

Peter, I. S. and Davidson, E. H. (2009). Genomic control of patterning. *Int. J. Dev. Biol.* **53**, 707-716.

Pollet, N., Muncke, N., Verbeek, B., Li, Y., Fenger, U., Delius, H. and Niehrs, C. (2005). An atlas of differential gene expression during early Xenopus embryogenesis. *Mech. Dev.* **122**, 365-439.

Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W. and Sandelin, A. (2010). JASPAR **2010**, the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105-D110.

Price, M. N., Huang, K. H., Arkin, A. P. and Alm, E. J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.* **15**, 809-819.

Quiring, R., Wittbrodt, B., Henrich, T., Ramialison, M., Burgtorf, C., Lehrach, H. and Wittbrodt, J. (2004). Large-scale expression screening by automated whole-mount in situ hybridization. *Mech. Dev.* **121**, 971-976.

Ramialison, M., Bajoghli, B., Aghaallaei, N., Ettwiller, L., Gaudan, S., Wittbrodt, B., Czerny, T. and Wittbrodt, J. (2008). Rapid identification of PAX2/5/8 direct downstream targets in the otic vesicle by combinatorial use of bioinformatics tools. *Genome Biol.* **9**, R145.

Rembold, M., Lahiri, K., Foulkes, N. S. and Wittbrodt, J. (2006). Transgenesis in fish: efficient selection of transgenic fish by co-injection with a fluorescent reporter construct. *Nat. Protoc.* **1**, 1133-1139.

Richardson, L., Venkataraman, S., Stevenson, P., Yang, Y., Burton, N., Rao, J., Fisher, M., Baldock, R. A., Davidson, D. R. and Christiansen, J. H. (2010). EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res.* **38**, D703-D709.

Rudnicki, M. A., Schnegelsberg, P. N., Stead, R. H., Braun, T., Arnold, H. H. and Jaenisch, R. (1993). MyoD or Myf-5 is required for the formation of skeletal muscle. *Cell* **75**, 1351-1359.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M. et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374-378.

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., Li, J., Thiagarajan, M., White, J. A. and Quackenbush, J. (2006). TM4 microarray software suite. *Methods Enzymol.* **411**, 134-193.

Saijou, E., Fujiwara, T., Suzaki, T., Inoue, K. and Sakamoto, H. (2004). RBD-1, a nucleolar RNA-binding protein, is essential for Caenorhabditis elegans early development through 18S ribosomal RNA processing. *Nucleic Acids Res.* **32**, 1028-1036.

Spitz, F., Gonzalez, F. and Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**, 405-417.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249-255.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S. E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S. E. et al. (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.* **3**, RESEARCH0088.

Uechi, T., Nakajima, Y., Nakao, A., Torihara, H., Chakraborty, A., Inoue, K. and Kenmochi, N. (2006). Ribosomal protein gene knockdown causes developmental defects in zebrafish. *PLoS ONE* **1**, e37.

Visel, A., Carson, J., Oldekamp, J., Warnecke, M., Jakubcakova, V., Zhou, X., Shaw, C. A., Alvarez-Bolado, G. and Eichele, G. (2007). Regulatory pathway analysis by high-throughput in situ hybridization. *PLoS Genet.* **3**, 1867-1883.

Vogel, J. H., von Heydebreck, A., Purmann, A. and Sperling, S. (2005). Chromosomal clustering of a human transcriptome reveals regulatory background. *BMC Bioinformatics* **6**, 230.

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276-287.

Wischnewski, J., Sölter, M., Chen, Y., Hollemann, T. and Pieler, T. (2000). Structure and expression of Xenopus karyopherin-beta3: definition of a novel synexpression group related to ribosome biogenesis. *Mech. Dev.* **95**, 245-248.

Wittbrodt, J., Shima, A. and Schartl, M. (2002). Medaka – a model organism from the Far East. *Nat. Rev. Genet.* **3**, 53-64.

Wong, L., Weadick, C. J., Kuo, C., Chang, B. S. and Tropepe, V. (2010). Duplicate dmbx1 genes regulate progenitor cell cycle and differentiation during zebrafish midbrain and retinal development. *BMC Dev. Biol.* **10**, 100.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K. et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7.

You, H. and Mak, T. W. (2005). Crosstalk between p53 and FOXO transcription factors. *Cell Cycle* **4**, 37-38.

Zalik, S. E., Lewandowski, E., Kam, Z. and Geiger, B. (1999). Cell adhesion and the actin cytoskeleton of the enveloping layer in the zebrafish embryo during epiboly. *Biochem. Cell Biol.* **77**, 527-542.

Zhang, Y., Miki, T., Iwanaga, T., Koseki, Y., Okuno, M., Sunaga, Y., Ozaki, N., Yano, H., Koseki, H. and Seino, S. (2002). Identification, tissue expression, and functional characterization of Otx3, a novel member of the Otx family. *J. Biol. Chem.* **277**, 28065-28069.

Zhou, Y., Luoh, S. M., Zhang, Y., Watanabe, C., Wu, T. D., Ostland, M., Wood, W. I. and Zhang, Z. (2003). Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res.* **63**, 5781-5784.

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. and Furlong, E. E. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65-70.