Development 138, 879-884 (2011) doi:10.1242/dev.055996

Minor change, major difference: divergent functions of highly conserved cis-regulatory elements subsequent to whole genome duplication events

Debbie K. Goode[1], Heather A. Callaway[2], Gustavo A. Cerda[1], Katharine E. Lewis[1,3] and Greg Elgar[4,*]

SUMMARY

Within the vertebrate lineage, a high proportion of duplicate genes have been retained after whole genome duplication (WGD) events. It has been proposed that many of these duplicate genes became indispensable because the ancestral gene function was divided between them. In addition, novel functions may have evolved, owing to changes in cis-regulatory elements. Functional analysis of the *PAX2/5/8* gene subfamily appears to support at least the first part of this hypothesis. The collective role of these genes has been widely retained, but sub-functions have been differentially partitioned between the genes in different vertebrates. Conserved non-coding elements (CNEs) represent an interesting and readily identifiable class of putative cis-regulatory elements that have been conserved from fish to mammals, an evolutionary distance of 450 million years. Within the *PAX2/5/8* gene subfamily, *PAX2* is associated with the highest number of CNEs. An additional WGD experienced in the teleost lineage led to two copies of *pax2*, each of which retained a large proportion of these CNEs. Using a reporter gene assay in zebrafish embryos, we have exploited this rich collection of regulatory elements in order to determine whether duplicate CNEs have evolved different functions. Remarkably, we find that even highly conserved sequences exhibit more functional differences than similarities. We also discover that short flanking sequences can have a profound impact on CNE function. Therefore, if CNEs are to be used as candidate enhancers for transgenic studies or for multi-species comparative analyses, it is paramount that the CNEs are accurately delineated.

KEY WORDS: Cis-regulatory elements, DDC model, Enhancer analyses, Evolution, Transcriptional regulation, Whole genome duplication events

INTRODUCTION

It is widely accepted that during vertebrate evolution two whole genome duplication (WGD) events occurred, followed by another in the lineage leading to teleosts (Amores et al., 1998; Dehal and Boore, 2005; Holland et al., 1994; Ohno et al., 1968; Taylor et al., 2003; Wittbrodt and Schartl, 1998). These events coincided with a rapid expansion in organismal complexity particularly in teleosts, a lineage that constitutes half of extant vertebrates. The persistence of many duplicate genes after these WGD events forms the basis of the duplication-degeneration-complementation (DDC) model of gene evolution (Force et al., 1999). The DDC model predicts that at least some aspects of ancestral gene function are sub-partitioned between duplicate genes in a complementary manner, such that each copy remains indispensable. At the same time, redundancy of associated cis-regulatory elements may increase the 'evolvability' of these sequences (Jimenez-Delgado et al., 2009) and their potential to direct novel gene sub-functions. These ideas have led to the hypothesis that, rather than changes to protein-coding sequences, divergence in transcriptional regulation is the main driving force behind innovations in the vertebrate body plan (Aburomia et al., 2003; Levine and Tjian, 2003). Here, we explore these concepts using the *PAX2/5/8* gene subfamily as a model and, in the process, discover key considerations for CNE and enhancer studies.

The highly related vertebrate *PAX2*, *PAX5* and *PAX8* genes derive from the two pan-vertebrate WGD events while the more recent teleost specific WGD resulted in two co-orthologous *pax2* genes: *pax2a* and *pax2b* in zebrafish, and *pax2.1* and *pax2.2* in other teleosts (Pfeffer et al., 1998; Wada et al., 1998). The current data suggest that these genes have evolved in a manner consistent with the DDC model. In all vertebrates examined so far, *PAX2*, *PAX5* and *PAX8* collectively have important functions in the development of the CNS, eye, ear, kidney and thyroid, but the roles of individual genes have diverged both within the subfamily and across species (reviewed by Goode and Elgar, 2009). Bouchard and colleagues provided tangible evidence that *cis*-regulatory elements may be responsible for at least some of the functional divergence of these genes. They demonstrated that the insertion of *Pax5* cDNA into the *Pax2* locus is able to rescue *Pax2*-mutant phenotypes in mouse, even in domains where *Pax5* is not normally expressed (Bouchard et al., 2000). Therefore, given the correct regulatory environment, mouse *Pax5* at least is capable of substituting for *Pax2*.

In this paper, we investigate an interesting class of putative cis-regulatory elements consisting of non-coding sequences that are highly conserved between Fugu and humans [an evolutionary distance of around 450 million years (Sandelin et al., 2004; Woolfe et al., 2005)]. These conserved non-coding elements (CNEs) cluster around genes that are involved in transcriptional and developmental regulation and many exhibit in vivo enhancer activity in model

[1]Department of Physiology Development & Neuroscience, Anatomy Building, Downing Street, Cambridge CB2 3DY, UK. [2]Queen Mary, University of London, Mile End Road, London E1 4NS, UK. [3]Syracuse University, Department of Biology, 114 Life Sciences Complex, Syracuse, NY 13244, USA. [4]The MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UK.

*Author for correspondence (gelgar@nimr.mrc.ac.uk)

Accepted 20 December 2010

organisms, including zebrafish (de la Calle-Mustienes et al., 2005; Shin et al., 2005; Woolfe et al., 2005), mouse (Pennacchio et al., 2006), frog (de la Calle-Mustienes et al., 2005) and chick (Sabherwal et al., 2007). Additional evidence that at least some of these CNEs regulate gene expression has been provided by the identification of point mutations (Lettice et al., 2003; Benko et al., 2009), deletions (D'Haene et al., 2009; Sabherwal et al., 2007) and translocations (Gill et al., 2009) of individual CNE sequences that produce mutant phenotypes owing to dysregulation of the associated protein-coding region.

Previous analyses (Woolfe and Elgar, 2007) (http://condor.nimr.mrc.ac.uk/) have shown that the vertebrate *PAX2* gene is associated with a large number of CNEs (around 60). Interestingly, many tetrapod *PAX2* CNEs have sequence homology to both teleost *pax2* loci, suggesting that a large proportion of CNEs have been retained in duplicate subsequent to the WGD event that occurred in the teleost lineage. Here, we have exploited the wealth of these CNE duplicates and analysed their sequences in relation to the single tetrapod CNE copies. Coupling this with intra-species comparative functional analyses has enabled us to assess their function with regard to the DDC model. Strikingly, our results show that most duplicate CNEs have differences in their enhancer activities and that even highly similar sequences can direct very different patterns of reporter gene expression.

MATERIALS AND METHODS

Bioinformatic analyses

CNEs associated with *PAX2/5/8* gene loci were originally identified from the CONDOR database (Woolfe et al., 2007) (http://condor.nimr.mrc.ac.uk/). Subsequently, sequences from multiple species were extracted from Ensembl (Hubbard et al., 2009) (http://www.ensembl.org/index.html). These were aligned using MLAGAN (http://lagan.stanford.edu/lagan_web/index.shtml) (Brudno et al., 2003), with a Vista graphical output (Mayor et al., 2000). At the time that this analysis was performed, zebrafish *pax2a/b* loci had assembly errors, so Fugu was used as the model organism for comparative genomics and functional analyses. ClustalW (Thompson et al., 1994) was used for the alignment of individual CNEs.

Sequence conservation indices were calculated as a product of the proportion of sequence overlap between human and Fugu CNEs, and the proportion of identical bases, i.e. (length of overlapping Fugu sequence/length of human CNE) × (number of identical bases/length of human CNE). These are reported in the text as n±s.d.

PCR design

Nineteen pairs of CNEs that are retained in duplicate (are associated with both Fugu *pax2* co-orthologues) were selected from intergenic and intronic regions of the loci. These range in size from 57 to 432 bp and their percentage of shared sequence identity ranges from 77-97%. Where possible, oligonucleotides were designed using Primer 3 software (Rozen and Skaletsky, 2000). Otherwise, in order to be as close as possible to the CNE sequence, they were designed by eye, maximising the criteria for optimal primer design (as stipulated in Primer 3). CNEs were amplified and purified as described previously (Woolfe et al., 2005). CNE and oligonucleotide sequences are provided in the supplementary information (see Table S1 in the supplementary material).

Functional assay in zebrafish embryos

Purified CNEs were co-injected together with a GFP reporter gene under control of a human β-globin minimal promoter as previously described (Woolfe et al., 2005). Micro-injections were performed in one- to four-cell zebrafish embryos (day 1). Embryos were screened for GFP-positive cells and scored on day 2 and day 3 as described previously (Woolfe et al., 2005). Schematic diagrams and numbers of embryos with GFP expression in each domain have been deposited in our online database (http://condor.nimr.mrc.ac.uk/). At least 25 embryos were scored for each CNE assayed.

We also performed control experiments to demonstrate that our results using this co-injection assay are comparable with a more conventional cloning strategy using Gateway Tol2 cloning vectors. We show these results in Fig. S5 and Table S2 in the supplementary material. Briefly, elements were either blunt-end cloned into a *Sma*I-digested 228 p5E-MCS vector (kindly provided by the Chien laboratory) (Kwan et al., 2007) or cloned into pENTR^{TM}5′-TOPO TA (Invitrogen #K591-20), after adding adenine overhangs (10-minute extension at 72°C using NEB Taq DNA polymerase, #M0273). A one-way LR reaction using LR Clonase II Plus (Invitrogen #12538-120) was then used to clone into R4-L1 basEGFPpA Tol2, a vector containing a carp β-actin minimal promoter and modified for efficient single 5′ entry. (This vector was created using sequences from 353-pENTRbasEgfp and 426-pDest Tol2pA, kindly provided by the Lawson lab) (Villefranc et al., 2007). Injection mix (5 pl of 50 ng/μl DNA, 25 ng/μl of transposase mRNA) was injected into one-cell embryos. The Transposase mRNA was synthesised using Ambion mMESSAGE mMACHINE (Invitrogen #AM1340M).

RESULTS AND DISCUSSION

The distribution of CNEs around the *PAX2/5/8* genes

Following WGD events, CNEs are often asymmetrically partitioned around gene duplicates (Woolfe and Elgar, 2007). The *PAX2/5/8* genes are no exception as they are, respectively, associated with around 60, 16 and two CNEs across all representative gnathostome groups (http://condor.nimr.mrc.ac.uk/) (Fig. 1; see Figs S1 and S2 in the supplementary material). By contrast, the teleost *pax2* co-orthologues have retained a similar number of CNEs, about two-thirds of which exist in duplicate (Woolfe and Elgar, 2007) (Fig. 1). Of these duplicate CNEs, four share sequence homology with *pax5* CNEs and another shares sequence homology with a *pax8* CNE (Fig. 1; see Fig. S3 in the supplementary material). In all cases, duplicate CNEs are in a similar position relative to the protein-coding region of the gene. There are no CNEs retained between all three paralogues and none are retained between *pax5* and *pax8* (Fig. 1; see Figs S1 and S2 in the supplementary material). Owing to the abundance of CNE duplicates retained between the teleost *pax2* co-orthologues and to the fact than we can compare them to the single copy tetrapod CNEs, we have used this dataset in order to explore CNE functionality subsequent to a WGD event.

A majority of *pax2* CNEs exhibit enhancer activity

Using an in vivo co-injection reporter gene assay in zebrafish embryos, we tested 19 pairs of *pax2* elements that are retained in duplicate between Fugu *pax2* co-orthologues. Remarkably, over 80% (31/38) of the elements are able to drive reporter gene expression in a tissue-specific and reproducible manner (Fig. 2), and in the majority of cases this strongly overlaps with endogenous gene expression. For example, most elements (27/38) drive expression in the hindbrain and spinal cord, and expression in thyroid and pronephros regions is also frequently observed (24 and 21 elements, respectively).

Duplicate *pax2* CNEs have diverged in function

Strikingly, we observed more differences than similarities between the expression profiles of duplicate CNEs. Most obviously, there are four pairs of CNEs (4, 10, 11 and 17) in which one CNE is able to drive reporter gene expression (interestingly always the *pax2.2* CNE), while the other shows little or no GFP expression (Figs 2 and 3).

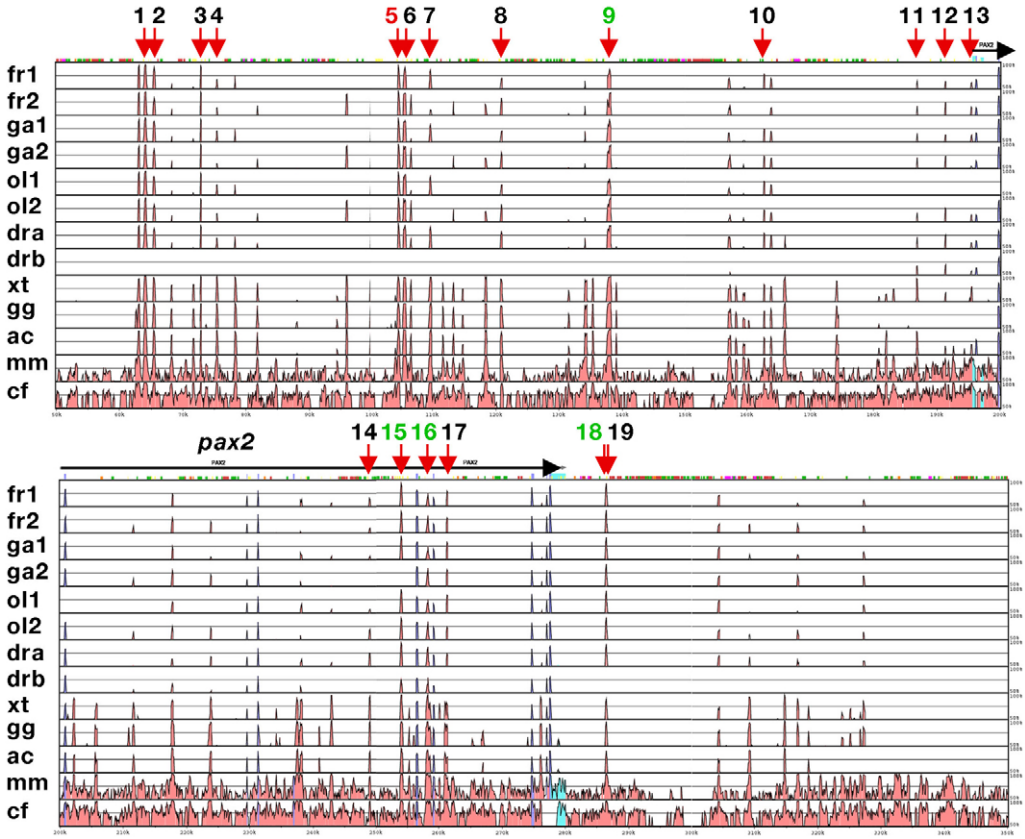

Fig. 1. Vista graphical output of a multi-species MLAGAN alignment using human sequence as a baseline. Peaks indicate sequence homology of pair-wise alignments between human and species indicated on the left (ac, anole lizard; cf, dog; dr, zebrafish; fr, Fugu; ga, stickleback; gg, chicken; mm, mouse; ol, medaka; xt, frog). Y-axis indicates percentage sequence identity. Blue peaks represent protein-coding regions; pale blue indicates the UTR; pink indicates CNEs. Red arrows indicate the 19 pairs of elements selected for this study. Coloured numbers indicate CNEs retained near *pax5* (green) or *pax8* (red) loci. Teleost species are labelled 1 or 2 for *pax2.1* and *pax2.2*, respectively (a and b in zebrafish).

The CNE pair with the most similar expression profile (15) drives expression predominantly in muscle but even here there are noticeable differences between individual CNEs. As well as spatial differences (the *pax2.1* element also drives expression in the CNS), there are temporal differences. Compared with day 2, there is a fourfold increase (39% versus 9%) in the percentage of GFP-positive embryos on day 3 for the *pax2.1* CNE, whereas this is less pronounced for the *pax2.2* CNE (45% versus 34%; Fig. 2).

In terms of spatial expression, the most dramatic difference is observed with CNE pair 1. In this case, the *pax2.1* CNE activates GFP in only a few regions, with most expression in the notochord and some expression in muscle and fin (Figs 2 and 3). By contrast, the *pax2.2* CNE drives GFP expression in a highly complex pattern in virtually all of the domains scored in this assay, except the notochord and the fin. This is an extraordinary result given the high similarity in sequence of these CNEs, both in terms of overlap (see Fig. S3 in the supplementary material) and sequence identity (~90%).

Comparison of duplicate CNE sequences

These striking differences in expression profiles led us to compare carefully the two sequences that we were using to test each pair of duplicate CNEs. In some cases (e.g. 8) the teleost CNEs overlap asymmetrically with the human sequence, and others (e.g. 2 and 7) are of unequal length (see Fig. S3 in the supplementary material). However, over one-third (37%) differ in length by less than 10 bp and align to the equivalent region of the single human *PAX2* locus (see Fig. S1 in the supplementary material). Although almost half (47%) of the *pax2.2* CNEs are shorter than the *pax2.1* counterparts, their average conservation indices are identical [respectively, 0.805±0.077 (s.d.) and 0.795±0.068 (s.d.)] suggesting a similar rate of divergence. In two cases (pairs 5 and 9) one CNE duplicate (*pax2.1* and *pax2.2*, respectively) contains an internal gap in conservation, suggesting a fragmentation of the CNE.

Non-conserved flanking sequences influence enhancer activity

Intriguingly, as mentioned above, the pair 1 elements have highly similar overlapping sequences, and yet, their expression profiles are dramatically different. The PCR products for this CNE pair included very short additional sequences (50 bp 5′ and 3′, respectively, for the *pax2.1* and *pax2.2* elements). In order to rule out any potential influence from these sequences, we repeated our assay using PCRs derived from sub-optimal primers located at the boundaries of the CNE sequence.

Interestingly, we see a striking difference in the expression profiles of these more tightly defined CNE sequences, compared with the original elements (Fig. 3). Expression driven by the new *pax2.1* element is more complex and includes domains (e.g. cardiovascular) that overlap with those produced by the *pax2.2* element. In contrast to the original element, the new *pax2.2* element drives expression in fin, more expression in muscle, less expression in CNS and no expression in sensory organs. However, although the new *pax2.1*/*pax2.2* expression profiles are more similar than the original ones, they are still not identical. Unlike the *pax2.2* element, the *pax2.1* element does not drive expression in CNS but does drive expression in the otic vesicle. Likewise, the *pax2.2* expression profile has lower expression in skin and fin, higher cardiovascular expression and lacks notochord expression. Therefore, even highly similar sequences (with 88% sequence identity) can be functionally divergent.

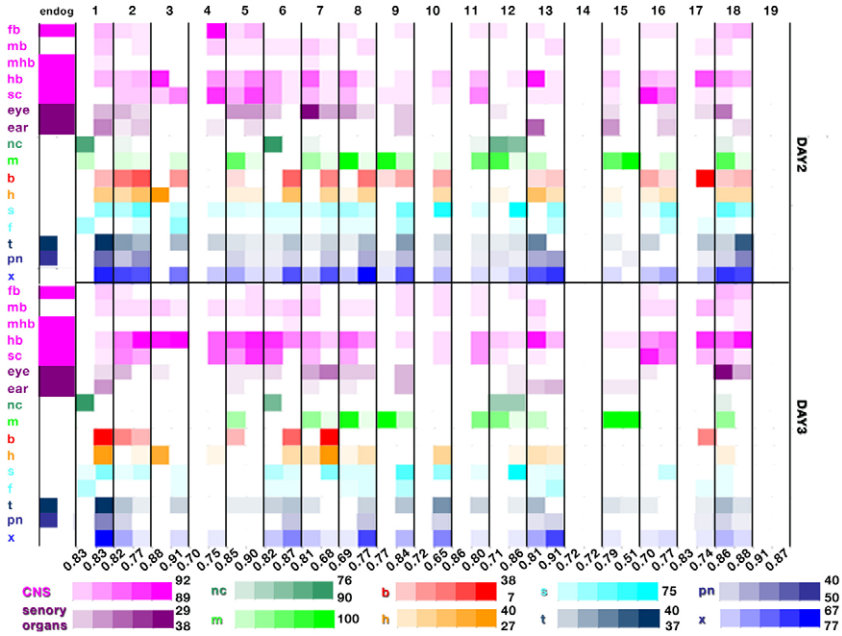

Fig. 2. Heat map depicting results of reporter gene assays to test function of CNEs retained in duplicate between two Fugu *pax2* co-orthologues. Results of each duplicate CNE pair are shown in vertical dual-panels, *pax2.1* on the left and *pax2.2* on the right. Expression was scored for each embryonic domain as indicated on the left and colour coded according to the key below. Colour intensity reflects percentage of GFP-positive embryos with expression in each domain. The maximum percentage is shown to the right of each key (day 2 top, day 3 bottom as shown in the heat map). Endogenous expression domains are indicated in the first panel on the left. The conservation index between the single human and each Fugu CNE is given at the bottom of each panel. b, blood; endog, endogenous expression; f, fin; fb, forebrain; h, heart; hb, hindbrain; m, muscle; mb, midbrain; mhb, mid-hindbrain boundary; nc, notochord; pn, pronephros region; s, skin; sc, spinal cord; t, thyroid region; x, other unidentified.

These remarkable results led us to re-analyse another pair of elements, pair 4, which have substantial differences in their expression profiles (Figs 2 and 3). Whereas the *pax2.2* element shows strong enhancer activity (over 30% of embryos have GFP expression), the *pax2.1* element drives very little expression (0.8% and 2.7% of embryos express GFP on day 2 and day 3, respectively). In this case, the *pax2.2* element extends beyond the defined CNE region but shows good overlap and alignment with the *pax2.1* element (see Fig. S4 in the supplementary material). However, the *pax2.1* sequence incorporates an additional 50 bp 5′ of this alignment (see Fig. S3 in the supplementary material). Therefore, we re-assayed this element eliminating this extra sequence.

This new element has slightly more enhancer activity than the original one (21/512 embryos (4%) expressed GFP). This expression occurred in similar domains to the *pax2.2* element (in the hindbrain and spinal cord), but also elsewhere (the tectum and forebrain; Fig. 3). However, on day 3 we could detect GFP in only five out of 421 embryos (1.2%). Therefore, even this more tightly defined CNE sequence only has very weak enhancer activity, which is still significantly different from its duplicate *pax2.2* CNE.

Isolated cis-regulatory sequences drive expression in both endogenous and ectopic domains

Our co-injection assay consistently shows that the majority of CNEs are able to drive reporter gene expression in a tissue-specific and reproducible manner. Our results are consistent with results that we obtained using more conventional Tol2 cloning strategies and they are apparently independent of the basal promoter used (see Fig. S5 and Table S2 in the supplementary material). In most cases, GFP expression recapitulates endogenous gene expression domains, with most CNEs driving expression in hindbrain and spinal cord, many driving expression in thyroid and pronephros regions, and about half driving expression in sensory organs. This strongly suggests that these sequences normally regulate *Pax2* expression. However, many of the CNE expression profiles are complex, and frequently include ectopic domains. This may be because we are testing these elements as isolated sequences outside the influence of the genomic environment.

Ectopic expression is a common, but little discussed, occurrence in assays using isolated cis-regulatory sequences and transgenic lines constructed using subpopulations of cis-regulatory elements (see Goode and Elgar, 2009). For example, mouse (Ohyama and Groves, 2004; Rowitch et al., 1999) and zebrafish (Picker et al., 2002) *pax2* transgenic lines recapitulate many aspects of endogenous *Pax2* expression, but they also have ectopic expression and expression within endogenous domains is not always temporally correct. Strikingly, none of these transgenic lines exhibits expression in the eye, even though an optic stalk enhancer (Schwarz et al., 2000) is embedded within the upstream sequences used to generate these lines. The potential activity of an element can, therefore, be missed if it is not analysed in isolation. However, our results show that CNEs often have the ability to drive reporter gene expression that both overlaps with and occurs outside endogenous domains. Notably, regardless of methodology (see Fig. S5 and Table S2 in the supplementary material) even the ‘ectopic’ expression is reproducible across a large number of embryos. This indicates the potential enhancer properties of CNEs, which can be seen when they are tested in isolation but which are presumably latent (or repressed in some way) when these sequences are in their normal context within the genome. Cumulative data such as these are invaluable for determining the underlying sequence language that can regulate gene expression and also for identifying potential sequences that might be responsible for anomalous gene expression when the gene regulatory landscape is perturbed by mutations or translocations.

Highly similar sequences can have divergent functions

Strikingly, our results show that even highly conserved duplicate CNE pairs can have different expression profiles. This illustrates the relative ease with which functionality can evolve with little sequence change. Like the single tetrapod *PAX2* gene, the teleost co-orthologues have complex expression patterns in the CNS, eye and ear. However, although their expression overlaps in these domains, there are temporal-spatial differences. In addition, only *pax2a* is expressed in the kidney and thyroid (Goode and Elgar, 2009; Pfeffer et al., 1998). This functional partitioning is consistent with the DDC model and experimental evidence shows that the co-

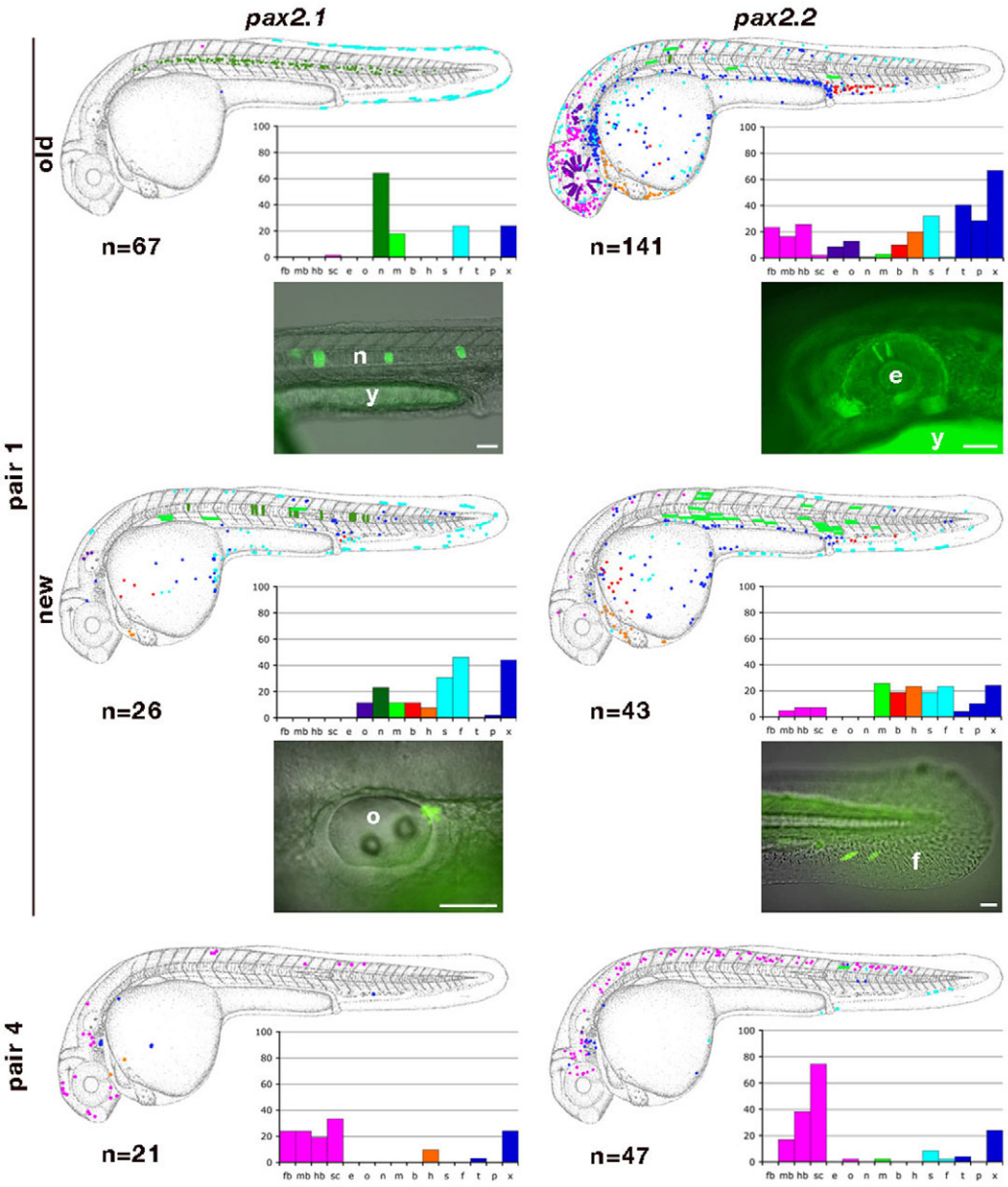

Fig. 3. Schematic diagrams of expression profiles derived from assaying *pax2* element pairs 1 and 4. Expression in each domain is colour coded according to the key in Fig. 2 and mapped onto camera lucida drawings of day 2 and day 3 zebrafish embryos (day 2 is shown here). Results are overlaid from multiple embryos. *n* indicates the number of embryos analyzed. The percentage of GFP-positive embryos with expression in each domain (*y*-axis) is indicated in the respective bar charts. For pair 1, 'old' indicates results from our original assay and 'new' indicates results from our newer assay without the flanking sequences. Live images of day 3 embryos are also shown for these assays, with expression in the notochord (*pax2.1* old), eye (*pax2.2* old), otic vesicle (*pax2.1* new) and fin (*pax2.2* new). Expression in the eye is shown as a fluorescent image, whereas the rest are shown as merged fluorescent and bright-field images. b, blood; e, eye; f, fin; fb, forebrain; h, heart; hb, hindbrain; m, muscle; mb, midbrain; n, notochord; o, otic vesicle; p, pronephric region; s, skin; sc, spinal cord; t, thyroid region; x, other, unclassified; y, yolk. Scale bars: 50 μm.

orthologues are able to functionally substitute for one another in overlapping expression domains (reviewed by Goode and Elgar, 2009). The retention of both duplicates therefore adds a robustness to some *Pax2* functions while allowing others to diverge between the duplicated genes. From our analyses of *Pax2* CNEs, we can readily appreciate how subtle sequence changes within these elements could rapidly expand the already complex range of *pax2* functions, thus influencing the evolving animal body plan.

Given our results, it is crucial that assays of CNEs should use sequences that are as accurately delineated as possible. Even our initial stringent primer design strategy included short additional sequences, and yet as we have shown, these can dramatically influence the activity of CNEs. This is a crucially important finding, given the prevalence with which CNEs are now used to try and construct transgenic animals and/or drive expression in specific cells and tissues.

Importance of our results for cross-species comparisons of CNEs

A powerful way of assessing sub-functions embedded within duplicate CNEs is to use phylogenetic and functional analyses of multiple vertebrate sequences. Here, we have shown the potential behind comparing duplicate teleost CNEs with single-copy tetrapod sequences. Extending such analyses should enable us to identify which base changes are permissive during the evolution of these sequences and how subtle sequence differences affect the regulatory ability of CNEs. Multi-species comparisons also allow us to delineate putative cis-regulatory elements in terms of the boundary of sequence conservation. Given that as little as 50 bp of non-conserved flanking sequence can dramatically influence CNE enhancer activity, it is obviously important to identify the limits of CNE sequences as accurately as possible. Our results strongly suggest that once a consensus has been reached, these boundaries need to be strictly adhered to when analysing CNEs if we are to be able to compare functional data from related genes and different species. Only then will we be able to confidently interpret functional data in order to generate an evolutionary profile of these extraordinary sequences.

Acknowledgements

We thank Julie Cooke, Sarah Smith, Phil Snell and Phil North for contributions to early experimental work. This work was supported by an MRC project grant (72504) awarded to G.E., by a Wellcome Trust project grant awarded to K.E.L. (Ref 079971) and by a Royal Society University Research Fellowship, also awarded to K.E.L. Deposited in PMC for release after 6 months.

Competing interests statement

The authors declare no competing financial interests.

Supplementary material

Supplementary material for this article is available at
http://dev.biologists.org/lookup/suppl/doi:10.1242/dev.055996/-/DC1

References

Aburomia, R., Khaner, O. and Sidow, A. (2003). Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. *J. Struct. Funct. Genomics* **3**, 45-52.

Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L. et al. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**, 1711-1714.

Benko, S., Fantes, J. A., Amiel, J., Kleinjan, D. J., Thomas, S., Ramsay, J., Jamshidi, N., Essafi, A., Heaney, S., Gordon, C. T. et al. (2009). Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat. Genet.* **41**, 359-364.

Bouchard, M., Pfeffer, P. and Busslinger, M. (2000). Functional equivalence of the transcription factors Pax2 and Pax5 in mouse development. *Development* **127**, 3703-3713.

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A. and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721-731.

D'Haene, B., Attanasio, C., Beysen, D., Dostie, J., Lemire, E., Bouchard, P., Field, M., Jones, K., Lorenz, B., Menten, B. et al. (2009). Disease-causing 7.4 kb cis-regulatory deletion disrupting conserved non-coding sequences and their interaction with the FOXL2 promotor: implications for mutation screening. *PLoS Genet.* **5**, e1000522.

de la Calle-Mustienes, E., Feijoo, C. G., Manzanares, M., Tena, J. J., Rodriguez-Seguel, E., Letizia, A., Allende, M. L. and Gomez-Skarmeta, J. L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* **15**, 1061-1072.

Dehal, P. and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314.

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

Gill, H. K., Parsons, S. R., Spalluto, C., Davies, A. F., Knorz, V. J., Burlinson, C. E., Ng, B. L., Carter, N. P., Ogilvie, C. M., Wilson, D. I. et al. (2009). Separation of the PROX1 gene from upstream conserved elements in a complex inversion/translocation patient with hypoplastic left heart. *Eur. J. Hum. Genet.* **17**, 1423-1431.

Goode, D. K. and Elgar, G. (2009). The PAX258 gene subfamily: a comparative perspective. *Dev. Dyn.* **238**, 2951-2974.

Holland, P. W., Garcia-Fernandez, J., Williams, N. A. and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Development* **Suppl**, 125-133.

Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L. et al. (2009). Ensembl 2009. *Nucleic Acids Res.* **37**, D690-D697.

Jimenez-Delgado, S., Pascual-Anaya, J. and Garcia-Fernandez, J. (2009). Implications of duplicated cis-regulatory elements in the evolution of metazoans: the DDI model or how simplicity begets novelty. *Brief. Funct. Genomic. Proteomic.* **8**, 266-275.

Kwan, K. M., Fujimoto, E., Grabher, C., Mangum, B. D., Hardy, M. E., Campbell, D. S., Parant, J. M., Yost, H. J., Kanki, J. P. and Chien, C. B. (2007). The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs. *Dev. Dyn.* **236**, 3088-3099.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E. and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725-1735.

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature* **424**, 147-151.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S. and Dubchak, I. (2000). VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046-1047.

Ohno, S., Wolf, U. and Atkin, N. B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169-187.

Ohyama, T. and Groves, A. K. (2004). Generation of Pax2-Cre mice by modification of a Pax2 bacterial artificial chromosome. *Genesis* **38**, 195-199.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D. et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499-502.

Pfeffer, P. L., Gerster, T., Lun, K., Brand, M. and Busslinger, M. (1998). Characterization of three novel members of the zebrafish Pax2/5/8 family: dependency of Pax5 and Pax8 expression on the Pax2.1 (noi) function. *Development* **125**, 3063-3074.

Picker, A., Scholpp, S., Bohli, H., Takeda, H. and Brand, M. (2002). A novel positive transcriptional feedback loop in midbrain-hindbrain boundary development is revealed through analysis of the zebrafish pax2.1 promoter in transgenic lines. *Development* **129**, 3227-3239.

Rowitch, D. H., Kispert, A. and McMahon, A. P. (1999). Pax-2 regulatory sequences that direct transgene expression in the developing neural plate and external granule cell layer of the cerebellum. *Dev. Brain Res.* **117**, 99-108.

Rozen, S. and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365-386.

Sabherwal, N., Bangs, F., Roth, R., Weiss, B., Jantz, K., Tiecke, E., Hinkel, G. K., Spaich, C., Hauffa, B. P., van der Kamp, H. et al. (2007). Long-range conserved non-coding SHOX sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. *Hum. Mol. Genet.* **16**, 210-222.

Sandelin, A., Bailey, P., Bruce, S., Engstrom, P. G., Klos, J. M., Wasserman, W. W., Ericson, J. and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99.

Schwarz, M., Cecconi, F., Bernier, G., Andrejewski, N., Kammandel, B., Wagner, M. and Gruss, P. (2000). Spatial specification of mammalian eye territories by reciprocal transcriptional repression of Pax2 and Pax6. *Development* **127**, 4325-4334.

Shin, J. T., Priest, J. R., Ovcharenko, I., Ronco, A., Moore, R. K., Burns, C. G. and MacRae, C. A. (2005). Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res.* **33**, 5437-5445.

Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. and Van de Peer, Y. (2003). Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* **13**, 382-390.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.

Villefranc, J. A., Amigo, J. and Lawson, N. D. (2007). Gateway compatible vectors for analysis of gene function in the zebrafish. *Dev. Dyn.* **236**, 3077-3087.

Wada, H., Saiga, H., Satoh, N. and Holland, P. W. (1998). Tripartite organization of the ancestral chordate brain and the antiquity of placodes: insights from ascidian Pax-2/5/8, Hox and Otx genes. *Development* **125**, 1113-1122.

Wittbrodt, J. M. A. and Schartl, M. (1998). More genes in fish? *BioEssays* **20**, 511-515.

Woolfe, A. and Elgar, G. (2007). Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. *Genome Biol.* **8**, R53.

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.

Woolfe, A., Goode, D. K., Cooke, J., Callaway, H., Smith, S., Snell, P., McEwen, G. K. and Elgar, G. (2007). CONDOR: a database resource of developmentally associated conserved non-coding elements. *BMC Dev. Biol.* **7**, 100.

Table S1. Primer sequences used for the PCR amplification of *pax2* CNEs

CNE ID	Gene	Forward primer	Reverse primer
CRCNE00000055old	*pax2.1*	CACTCACAGATTCCATCAAAGCC	AGAGAGTGAGTGACAAAGCAGG
CRCNE00000055new	*pax2.1*	AATGAAATTGAGACATGGTAG	AGAGAGTGAGTGACAAAGCAGG
CRCNE00000123old	*pax2.2*	GAATTAAATGGAGACACGGTGGC	CCATTCTCTCCGCATGAGGG
CRCNE00000123new	*pax2.2*	GAATTAAATGGAGACACGGTGGC	AAAGGAGAGGAAAAGCACGGG
CRCNE00000056	*pax2.1*	CTTTGTCACTCACTCTCTCCC	GATGTGACATACAACCACGCC
CRCNE00000125	*pax2.2*	GCACTGACAGGCAGATCCC	CCTTGATTAGCAGCAGCCTTCC
CRCNE00000059	*pax2.1*	CTCCCTGTGCTAATCCCTCATC	CAACACCAGCTCAGGGATGAAG
CRCNE00000127	*pax2.2*	AGGACAGAAAACCTAGCCTCAG	GCAGGCAATGATTAATGTCCCTC
CRCNE00000060old	*pax2.1*	CACAAGTGGTGAGTCTGAGC	GTGATTTGTTGCGGCGATGC
CRCNE00000060new	*pax2.1*	GAAGGCAATCTGTTCATTAAG	GTGATTTGTTGCGGCGATGC
CRCNE00000128	*pax2.2*	CTTGCTCATTACTAGCTTGCG	TGCTCACTGATTTGTTACCAC
CRCNE00000063-4	*pax2.1*	CCCTACAATCCATCTTGTGGAG	CCGTTTGGAGCCTGTTCCC
CRCNE00000133	*pax2.2*	CCATCCTCAACTCAGATAGTCC	CGTTTGGAGCTTGTACTCCCAG
CRCNE00000065	*pax2.1*	GCCTACAAACAGAAGCAGGACC	GGTGGATCACTCAACCGTGAC
CRCNE00000134	*pax2.2*	TAATATCTAGCCGATGCCCTG	GGAACGGCTCGCTTGTGC
CRCNE00000066	*pax2.1*	GTTCTGCCTCTGCATATTCC	TACCTGCAACACTAAAGATCC
CRCNE00000136	*pax2.2*	AGTGAAGTGAATAAGCCACAAC	TTTGCTTCTTGTTCTGTGACC
CRCNE00000068	*pax2.1*	CTAAGGTCGCCCTGTCGAGG	TCGCTTTATGGCCAGACCTTC
CRCNE00000140	*pax2.2*	CTTCCGGTCCACGACCTCC	CTAATGGCTATCTTGAAGTGC
CRCNE00000071	*pax2.1*	CATCCCTGCCCGATAATCTGG	CTCAGCCAAATGGCAGGCTTC
CRCNE00000145-7	*pax2.2*	GGGTAGGTTATTCAGTGAGCATC	CTCATAATTCCTCGCACGCCG
CRCNE00000078	*pax2.1*	GTGCATGACTGTCCCTCTC	GCATGAGACACTGTGGAGG
CRCNE00000151	*pax2.2*	ACGCCTCCAGAGGCAGTGG	CAAACGGAGCGACAGTGGGAC
CRCNE00000080	*pax2.1*	GGGGAAGATCAATGGAAAACAC	CTTGGGTCGCTTCTACACG
CRCNE00000154	*pax2.2*	CGAGGATCAATGGAAAACAGTTG	GCTGCACTGTACTGAACTCC
CRCNE00000081	*pax2.1*	GTAGTTCTGTGACACCAGGACG	GAGGGAGCTTCTGGTACAATATG
CRCNE00000155	*pax2.2*	ACAGAGGGAGAGGTGGG	TCTCACGCTCTGCCTCC
CRCNE00000082	*pax2.1*	CCCCAAAGGCTCATCATTTCCC	CTCGCATAGAGATGGATGACTTG
CRCNE00000156	*pax2.2*	CTCAGACCTCCTCATTTGGAC	CCCCTTTTGCTTATATGGATGAC
CRCNE00000089	*pax2.1*	CGCTCGTCCACACTGAATG	GTAGCACAAGAAGAAACTGGG
CRCNE00000164	*pax2.2*	CTAATGCTGCGGCACAGGC	GCTGTTGTCTTTGCTACTCAGG
CRCNE00000090	*pax2.1*	GCGAGTTGGGTTCATCCTCTG	CAACAGTGGATGAGGACTTTAGC
CRCNE00000165	*pax2.2*	GGTCGGGCTCATCCTCTG	CCTCTGTCGCTGCATTCCC
CRCNE00000091-2	*pax2.1*	CTTGCTCAAAGCCTGTAAATCC	CCTCCGCTCTTTGTGATTGC
CRCNE00000166-7	*pax2.2*	TACATACCACAGCCACACTTGC	ACCCTTCTCGATGCTCTTTGTG
CRCNE00000095	*pax2.1*	TTCTGCCATAAGCAAACCCTG	GGGTAAACAGAGGACGCC
CRCNE00000171	*pax2.2*	CAGCGATAAGCCTATCAGGG	CCAATCTCTGCCAATGAGCG
CRCNE00000099	*pax2.1*	GAACAGATGAGGCAACGAGG	GTTTGCCAAAGAGGGGCTAC
CRCNE00000174	*pax2.2*	GCACCAGCTCACTCCCAAC	CGAGCTACTTTGCTCTCTTTGC
CRCNE00000100	*pax2.1*	GAGCGTGGTGGAAGTTAGTCTG	CATGCCTTCGTCTATGACAGGG
CRCNE00000175	*pax2.2*	TACCAGGGGAGTGCAGTGG	CATGGCTTCCTCTATGACAGG
CRCNE00000735	*pax8*	AGCGGCAGAGAGGGTAAAAG	AGCTCAGCTGAAAGCCACAG

CONDOR database (http://condor.nimr.mrc.ac.uk/). Identification numbers of CNEs are given in the first column, followed by the associated Fugu *pax2* co-orthologue. Where CNEs were re-assayed, sequences used in the original and repeat assays are referred to as 'old' and 'new', respectively. Additional primer sequences are given for the pax8 element tested using the Tol-2 assay.

Table S2. Summary of the results from analysing *pax2/8* CNEs using either the co-injection assay or Tol2 cloning

		Expression	
CNE ID	Gene	Co-injection	Tol2
CRCNE00000063-4	*pax2.1*	CNS , eye, cardiovascular, blood, muscle, pronephric region	CNS, eye, ear, cardiovascular, blood, muscle, pronephric region
CRCNE00000133	*pax2.2*	Hindbrain, spinal cord, telencephalon, eye, cardiovascular, muscle	Hindbrain, spinal cord, telencephalon, eye, cardiovascular, muscle
CRCNE00000735	*pax8*	CNS	CNS, eye, ear
CRCNE00000100	*pax2.1*	No expression out of 235 screened	No expression on day 2, 10/292 with expression in heart and/or skin on day 3
CRCNE00000175	*pax2.2*	No expression out of 244 screened	No expression out of 525 screened

CONDOR identifiers are given in the first column followed by the gene name and description of expression. For GFP-positive elements, we selected CNEs with shared sequence homology to both *pax2* co-orthologues and *pax8* (the first three elements listed here). As illustrated in Fig. S5, expression is highly similar. None of the expression domains described for the co-injection assay differed from those observed in the Tol2 results. The other two elements showed no evidence of enhancer activity using the co-injection assay. This was corroborated by the Tol2 system, with only a low level of expression on day 3 in the case of element CRCNE00000175. This may be due to the higher level of transient expression (both specific and ectopic) usually observed with the Tol2 method.