Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints

Manuel Irimia, Juan J Tena, Maria Alexis, et al.

Genome Res. published online June 21, 2012
Access the most recent version at doi:10.1101/gr.139725.112

Extensive Conservation of Ancient Microsynteny across Metazoans due to *cis*-regulatory Constraints

Manuel Irimia [1,2,7], Juan J. Tena [3,6], Maria S. Alexis [1,6], Ana Fernandez-Miñan [3,6], Ignacio Maeso [4], Ozren Bogdanović [3], Elisa de la Calle-Mustienes [3], Scott W. Roy [1,5], José L. Gómez-Skarmeta [3] and Hunter B. Fraser [1]

1 - Department of Biology, Stanford University, Stanford, CA, USA.

2 - The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada.

3 - Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, Sevilla, Spain.

4 - Department of Zoology, University of Oxford, Oxford, UK.

5 - Present address: Department of Biology, San Francisco State University, San Francisco, CA, USA.

6 - These authors contributed equally to this work.

7- Corresponding author: Manuel Irimia, mirimia@gmail.com

The Donnelly Centre, University of Toronto

160 College Street, Room 908

Toronto, Ontario, Canada M5S 3E1

Lab 416-978-7150

Fax 416-946-5545

Running Title: Deep Conservation of Microsynteny across Metazoans

Key words: Genome evolution, transcriptional regulation, synteny, highly conserved non-coding regions, evolutionary constraints.

Abstract

The order of genes in eukaryotic genomes has generally been assumed to be neutral, since gene order is largely scrambled over evolutionary time. Only a handful of exceptional examples are known, typically involving deeply conserved clusters of tandemly duplicated genes (e.g. Hox genes and histones). Here we report the first systematic survey of microsynteny conservation across metazoans, utilizing 17 genome sequences. We identified nearly 600 pairs of unrelated genes that have remained tightly physically linked in diverse lineages across over 600 million years of evolution. Integrating sequence conservation, gene expression data, gene function, epigenetic marks, and other genomic features, we provide extensive evidence that many conserved ancient linkages involve (i) the coordinated transcription of neighboring genes, or (ii) Genomic Regulatory Blocks (GRBs) in which transcriptional enhancers controlling developmental genes are contained within nearby bystander genes. In addition, we generated ChIP-seq data for key histone modifications in zebrafish embryos, which provided further evidence of putative GRBs in embryonic development. Finally, using chromosome conformation capture (3C) assays and stable transgenic experiments, we demonstrate that enhancers within bystander genes drive the expression of genes such as *Otx* and *Islet*, critical regulators of central nervous system development across bilaterians. These results suggest that ancient genomic functional associations are far more common than previously thought – involving ~12% of the ancestral bilaterian genome – and that *cis*-regulatory constraints are crucial in determining metazoan genome architecture.

Introduction

The evolutionary and functional implications of high-order eukaryotic genome structures remain topics of contention, and have inspired some of the most ambitious evolutionary hypotheses of the genetic and genomic eras (e.g. (Doolittle 1978; Lynch 2007; Koonin, Wolf 2010)). Largely absent from these debates has been the question of the local ordering of genes themselves within a genome, generally reflecting the common assumption that gene position within the genome is mostly a secondary concern and/or is usually not constrained (Koonin, Wolf 2010).

However, a variety of studies suggest that the situation is not so simple. Gene order is non-random (Hurst, Pal, Lercher 2004; Oliver, Misteli 2005; Michalak 2008), with clustering of genes in some species according to metabolic pathways and/or similarity of expression (Lee, Sonnhammer 2003; Fukuoka, Inaoka, Kohane 2004; Hurst, Pal, Lercher 2004). Particularly interesting are the Genomic Regulatory Blocks (GRBs, (Becker, Lenhard 2007; Engstrom et al. 2007; Kikuta et al. 2007)), which usually consist of (i) a *trans-dev* gene encoding a key transcriptional regulator with a complex spatiotemporal expression pattern, often involved in embryonic development (Woolfe et al. 2005; Kikuta et al. 2007); and (ii) nearby functionally-unrelated bystander gene(s), which contain *cis*-regulatory sequences for the *trans-dev* gene within their introns (Figure 1B). Gene linkage is thus important in GRBs since breakage of the microsyntenic association would disrupt *trans-dev*-associated *cis*-regulatory functions.

Despite these instances of non-randomness, questions remain about the generality and strength of natural selection in maintaining gene order. First, similarity of expression between neighboring genes could be the result, rather than the cause, of their proximity (e.g., because of shared chromatin structure). Second, direct studies of selection have produced ambiguous results, with evidence for purifying selection on maintaining gene order in hemiascomycetous yeasts (Hurst, Williams, Pal 2002; Fischer et al. 2006; Poyatos, Hurst 2007), but not in *Drosophila* (Weber, Hurst 2011), even over short phylogenetic distances. Third, gene order is nearly

completely scrambled between humans and model invertebrates (Putnam et al. 2008; Denoeud et al. 2010). Furthermore, although whole genome analyses of slow-evolving species and ancestral karyotype reconstructions showed that the chromosome-scale organization (so-called macrosynteny) has been largely conserved since the last common metazoan ancestor, these same studies found almost no traces of preservation of the specific ancestral local gene order (microsynteny) across metazoans (Putnam et al. 2007; Putnam et al. 2008; Srivastava et al. 2008; Srivastava et al. 2010).

Even in cases where selection appears to conserve microsynteny, questions remain about the generality and duration of these forces over evolutionary time. First, known GRBs are generally restricted to vertebrates (Kikuta et al. 2007) or insects (Engstrom et al. 2007), with only three transphyletic GRBs described to date (Wang et al. 2007; Irimia et al. 2012; Maeso et al. 2012). Second, while tandemly-duplicated *Drosophila* genes tend to remain linked (Quijano et al. 2008), such duplicates usually reflect recent duplications (Thomas 2007; Irimia, Maeso, Garcia-Fernandez 2008; Baldwin, Marko, Nelson 2009). This suggests that these associations are short-lived, again with only a few known exceptions (e.g. *Hox* genes and histone clusters). Third, while co-expression of linked genes is well-established (most strikingly, by usage of bi-directional promoters between genes (Adachi, Lieber 2002)), few gene linkages are conserved between eukaryotic kingdoms (Davila-Lopez, Martinez-Guerra, Samuelsson 2010). As such, the general paradigm holds that neutral processes dominate microsynteny, particularly over long evolutionary times (Srivastava et al. 2008; Koonin, Wolf 2010).

We report the first genome-wide analysis of microsynteny conservation across metazoans, analyzing 17 species spanning 1,000 million years (MY) of evolution. We identified 795 groups of genes that are associated in four or more major animal taxa, 595 of which correspond to unrelated (non-paralogous) genes, which we term Conserved Ancestral Microsyntenic Pairs (CAMPs).

Multiple lines of evidence suggest conservation of gene expression co-regulation for some CAMPs, and of ancient GRBs involving key *trans-dev* genes for others.

Results

Identification of ancient conserved gene associations across metazoans

For each pair of neighboring genes in the genome of a given species, we assessed whether the two genes were also tightly genetically linked (with ≤4 intervening genes) in other lineages (see Methods for details). We identified hundreds of pairs in each genome that were conserved across at least four major lineages, ranging from 197 in the tunicate *Ciona intestinalis* to 842 in the cephalochordate *Branchiostoma floridae*, and 600 in human (Table S1). To test significance we randomized gene order within each chromosome/scaffold. Across 100 replicates per species, we found an average of 0.015 conserved pairs using the same criteria, a false discovery rate lower than 0.0002 for all species (Table S1).

We then merged the data for each species into a single dataset, and filtered for potential annotation errors (see File S1), yielding a final set of 795 unique groups of gene pairs (or ≥2-gene clusters)(Table S2). Due to high duplication rates, precise orthology and paralogy relationships could not be established for 89/795 groups. These included previously described examples, such as the histone clusters (Davila-Lopez, Martinez-Guerra, Samuelsson 2010) and cytochrome-P450 genes (Thomas 2007; Baldwin, Marko, Nelson 2009), as well as eight groups of clusters/pairs of important developmental genes (*Hox*, *Wnt*, *Hes*, *En*, *Irx*, *Six*, *Tbx*, and other homeobox genes).

Another 110 groups included paralogous gene pairs whose group orthology could be more confidently assigned by BLAST. Interestingly, using Bayesian phylogenetic inferences (see File S1 for details), we found evidence that at least 68% of these pairs (Table S3) likely arose by independent tandem duplications in the different lineages and are therefore not ancestrally linked (e.g. the homologs of the human gene *C16orf5*, Figure S1). This suggests a significant level of

recurrent evolution of tandem gene duplicates across metazoan genomes (Maeso, Roy, Irimia 2012). Alternatively, however, these patterns could also reflect events of gene conversion between ancestral duplicates within each species.

Deeply conserved phylogenetically unrelated gene pairs

We also identified 595 groups of non-paralogous gene pairs. These pairs ranged in degree of conservation, with 377 of them conserved in four out of the 11 studied major metazoan lineages, 153 in five, 46 in six, and 19 in seven or more. Since these gene pairs did not result from tandem duplications, it is very unlikely that they have become linked independently in different lineages, and therefore are likely to represent ancient gene associations. We refer to these gene pairs as Conserved Ancestral Microsyntenic Pairs (CAMPs).

To expand the phylogenetic coverage, we next assessed conservation of these 595 CAMPs in four additional phylogenetically key species, whose incomplete genome assemblies or annotations precluded their inclusion in the initial analyses: the hemichordate *Saccoglossus kowalevskii*, the tunicate *Oikopleura dioica*, and two outgroups, the sponge *Amphimedon queenslandica* and the unicellular opistokont *Capsaspora owczarzaki*. With this information, we applied parsimony to reconstruct the evolutionary history of CAMPs. At least 378 CAMPs were already present at the origin of Eumetazoans (all animals but sponges and placozoans), and 593 at the origin of Bilateria (two pairs were specific to Protostomes). We also reconstructed degree of CAMP disruption (fraction of ancestral CAMPs lost) on each branch. Notably, despite several potential sources of error (see File S1), branches with high rates of CAMP disruption correspond closely to branches previously shown to have high rates of other genomic changes (e.g. intron loss, gene duplications, genome rearrangements, nucleotide substitutions, etc. (Kent, Zahler 2000; Lynch, Conery 2000; Stein et al. 2003; Bourque et al. 2005; Bhutkar et al. 2008; Irimia, Roy 2008; Putnam et al. 2008; Denoeud et al. 2010))(Figure 1C). For example, generally slow-evolving

species such as amphioxus and the mollusk *Lottia gigantea* have retained 448-492 CAMPs, whereas fast-evolving lineages such as nematodes (12 CAMPs), flies (46) and *C. intestinalis* (54) have conserved far fewer (File S1). These results are also in full agreement with previous studies on lineage-specific losses of chromosomal-scale gene linkage (macrosysteny)(Putnam et al. 2007; Putnam et al. 2008; Srivastava et al. 2008; Denoeud et al. 2010; Srivastava et al. 2010; Lv, Havlak, Putnam 2011). These results thus indicate that CAMP evolution closely follows overarching (though still poorly understood) trends of genome evolution.

Functional causes for evolutionary conservation of ancient gene associations

We next sought independent tests that CAMPs have been specifically maintained by selection. We tested predictions made by each of the two known general hypotheses for functional roles of the conservation of microsynteny: coordinated transcriptional regulation (i.e. co-regulation), and association of the genes as a GRB (Figure 1).

- *Co-regulation of gene pairs*

Linked genes may share common *cis*-regulatory sequences, leading to coordinated gene expression (Figure 1A). In this case, potential chromosomal breakpoints in the intergenic region between such linked pairs will disrupt expression. We tested six predictions of this scenario: (i) enrichment of divergently transcribed genes (i.e., in a head-to-head, or 5'-5', orientation), due e.g. to a bidirectional promoter; (ii) preferential conservation of this 5'-5' orientation; (iii) short intergenic distances; (iv) correlated expression patterns; (v) few insulator elements, and (vi) highly conserved intergenic sequences (due to functional transcriptional elements).

We found that most CAMPs were organized in a 5'-5' orientation (prediction (i)). In humans, 54.8% of CAMPs showed this orientation, compared to 26.7% of the control set (i.e. all non-paralogous neighboring gene pairs without conserved microsynteny; $p = 1.37 \times 10^{-9}$, χ^2 test;

Figure 2A). A similar pattern was found for all species (reaching 68.6% in *L. gigantea*) except for flies and tunicates (which have similar 3'-3' and 5'-5' orientations, see below). 5'-5' orientations are far more conserved (prediction (ii)): orientation was conserved across all species sharing the CAMP for 51.4% of CAMPs with a 5'-5' in human, compared to 15-25% for other orientations (Figure 2B); indeed, 5'-5' CAMPs account for 75.8% of human CAMPs with fully conserved orientation. A similar result was observed for other species (again with the exceptions of flies and tunicates), and for highly-conserved CAMPs (≥5 lineages; data not shown).

Next, we assessed whether CAMPs have shorter intergenic distances (prediction (iii)), which may facilitate coordinated expression (Davila-Lopez, Martinez-Guerra, Samuelsson 2010). In humans, intergenic regions of CAMPs were 2.3-fold shorter than the control set (mean of 48 vs. 113 Kbp, $p=2.3x10^{-5}$, KS test). This pattern was observed for all orientations, although it was strongest for 5'-5' orientations (41 vs. 126 Kbp, $p=1.2x10^{-6}$). Moreover, 5'-5' CAMPs with conserved orientation had the shortest intergenic regions of all subsets (29 Kbp). Finally, 5'-5' CAMPs were twice as likely to have very short intergenic regions (<1Kbp, a signature of bidirectional promoters (Adachi, Lieber 2002)) than the control set (12.5% vs 6%, p=0.001, χ^2 test).

To study coexpression of CAMPs (prediction (iv)), we calculated gene expression correlations across 23,941 different human microarray experiments (conducted using the Affymetrix U133 Plus 2.0 array). Since neighboring genes are known to have higher coexpression than unlinked genes (Fukuoka, Inaoka, Kohane 2004; Hurst, Pal, Lercher 2004), we compared each CAMP to a control set of 100 phylogenetically unrelated adjacent gene pairs with the same orientation and similar intergenic distance (see Methods). CAMPs showed significantly higher coexpression levels (average Spearman correlation coefficient, r=0.30) than both the control sets ($r=0.21$, $p=8.7x10^{-6}$, KS test) and random gene pairs ($r=0.14$, $p=9.6x10^{-21}$). For each CAMP, we then ranked its coexpression value relative to its control set and also relative to 100 random gene pairs. The cumulative plot shows an excess of high ranks with respect to both the control and the

random set (Figure 2C, black and gray areas above the red line, respectively). This excess is mostly due to an enrichment in the most highly coexpressed decile (ranks 91 to 100, red arrow in Figure 2D). This pattern was observed for all orientations, but was strongest for 5'-5'gene pairs (data not shown). Finally, to assess whether this may be a general pattern across metazoans, we performed a similar analysis for *D. melanogaster*, using 1,909 published microarray experiments. These showed even stronger coexpression, with 81% of the pairs ranking in the top half of their 100 matched control pairs (50% expected), and 25% in the most highly coexpressed decile (10% expected; Figure S2). This excess of coexpression of CAMPs relative to other neighboring pairs is consistent with selection to maintain some CAMPs because of shared transcriptional regulatory elements.

We next investigated whether intergenic regions separating CAMPs show few insulators (prediction (v)), which may impose a barrier for coordinated expression. Using insulators defined by chromatin signatures in nine human cell lines (Ernst et al. 2011), we found that CAMPs had significantly fewer intergenic insulators than their matched controls (0.062 vs. 0.076 insulators/Kbp, $p=6.4x10^{-5}$, KS test). This was more evident for 5'-5' orientations (0.054 vs. 0.070 insulators/Kbp, $p=1.4x10^{-5}$) and for relatively short intergenic regions (< 50 Kbp, 0.047 vs. 0.073 insulators/Kbp, $p=2.9x10^{-4}$). By contrast, CAMPs showed a higher density of transcriptional enhancers (Ernst et al. 2011) in intergenic regions (0.180 vs 0.155 enhancers/Kbp, p = 0.022), in particular for highly coexpressed genes (see below). Finally, intergenic sequence conservation was also significantly higher in CAMPs than in non-conserved pairs (average mammalian PhastCons score of 0.18 vs. 0.14, $p=2.9x10^{-6}$), fulfilling prediction (vi).

In summary, coordination of gene expression is likely to explain the preservation of many CAMPs. One example involves the mitochondrial chaperonin genes *Hspe1* and *Hspd1*, which are found together in nearly all studied species, from humans to the unicellular holozoan *Capsaspora* (and also in fungi (Davila-Lopez, Martinez-Guerra, Samuelsson 2010)), in a conserved 5'-5' orientation. In humans, they have a very short intergenic distance (<1Kbp) and are known to share a

bidirectional promoter (Hansen et al. 2003). In agreement with this, both genes show high coexpression levels in our microarray analysis (r=0.81, rank=99). Some other strongly coexpressed pairs in humans include: *UBA3-ARL6IP5* (r=0.84, rank=100), *HNRNPA2B1-CBX3* (r=0.84, rank=100), *HADHA-HADHB* (r=0.79, rank=100), and *ZMYM3-NONO* (r=0.83, rank=100).

- *Genome Regulatory Blocks (GRBs)*

Another major cause of conserved gene linkage may be preservation of GRBs. The typical GRB comprises a *trans-dev* gene and one or more functionally unrelated bystander genes, whose introns harbor *cis*-regulatory sequences that act on the *trans-dev* gene promoter (Figure 1B). Thus separation of the genes in a GRB is likely to result in misregulation of the *trans-dev* gene. Several features of known GRBs are not expected from gene pairs with coordinated expression, yielding four specific predictions: (i) presence of a *trans*-dev gene and one or more non-*trans-dev* bystander genes*;* (ii) extensive intronic sequence in the bystander gene; (iii) evolutionary conservation of intronic sequence in the bystander gene, specifically in relatively-short Highly Conserved Non-coding Regions (HCNRs); and (iv) transcriptional enhancers (acting on the *trans-dev* gene) in bystander introns.

In order to test these predictions, we first identified *trans-dev* genes using Gene Ontology (see Methods). 29.4% of conserved CAMPs in human were composed of a *trans-dev* and a non-*trans-dev* gene, compared to 19.5% in the control set ($p=6.0\times10^{-4}$, χ^2 test). Similar patterns were found in all species, with even higher proportions of *trans-dev*-plus-non-developmental CAMPs in fast-evolving species such as tunicates and flies (54.6% and 42.9%, respectively). This is consistent with the lower fraction of 5'-5' (putatively coregulated) CAMPs in these lineages.

We next studied bystander gene's introns (prediction (ii)). We found higher intron number ($p=7\times10^{-4}$, KS test), and average intron length (twice as long; $p=1.7\times10^{-6}$) in bystander genes relative to controls (Figure 3A,B). Greater intron length mostly reflected a subset of very long

introns (a feature generally associated with the presence of HCNRs (Irimia et al. 2011)): bystander genes have an average of 2.7 introns longer than 10 Kbp, compared to 0.9-1.1 in the other gene sets ($p=2.3x10^{-8}$, Figure 3C).

We also found higher conservation intron sequences in *trans-dev* and bystander genes than in the respective control sets (prediction (iii); p= 0.0017 and $p=3.2x10^{-5}$, respectively, using mammalian phastCons scores (Siepel et al. 2005)). In contrast, CAMPs with no *trans-dev* gene showed no greater intronic conservation than their matched controls (Figure 3D). Because enhancers likely constitute only a small fraction of all intronic sequence, we expect an even stronger enrichment of HCNRs than overall sequence conservation. Consistent with this, putative bystander genes in CAMPs had six to ten-fold more HCNRs than in non-conserved pairs, both for ancient conserved noncoding elements (aCNEs, (Lee et al. 2011))(6.2 vs. 1.0 aCNEs/Mbp, $p=1.4x10^{-4}$, Figure 3E), and for VISTA HCNRs (Visel et al. 2007)(1.6 vs. 0.2 VISTAe/Mbp, p=0.036). Finally, following prediction (iv), CAMP's bystander introns had a higher density of functionally defined transcriptional enhancers (Ernst et al. 2011) than the control set (0.194 vs. 0.180 enhancers/Kbp, $p=1.5x10^{-7}$)(Figure 3F). These four lines of evidence thus suggest that a subset of CAMPs are likely GRBs.

Experimental evidence for GRBs in vertebrate development

The ancient GRBs we identified typically involve genes from well known developmental gene families, such as Fox, Fgf, Tbx, Sox, Smad, etc. (Table S4). Therefore, despite the significantly higher density of enhancers active in cell lines (Ernst et al. 2011), the major effect of bystander-contained enhancers is expected during embryonic development. In order to test this hypothesis, we used chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) for three key epigenetic marks in zebrafish embryos at 24 hours post-fertilization (hpf): Histone 3 Lysine 4 trimethylation (H3K4me3, marks active promoters), Histone 3 Lysine 4

monomethylation (H3K4me1, often marks active enhancers when not overlapping with H3K4me3) and Histone 3 Lysine 27 trimethylation (H3K27me3, often marks inactive promoters, and indicates genes with tissue-specific expression patterns in whole embryos)(Turner 2007; Akkers et al. 2009; Margueron, Reinberg 2010). We searched for conservation of the 593 bilaterian CAMPs in the zebrafish genome, finding 260 conserved pairs of 205 unique groups, of which 29.2% included one *trans-dev* gene. Mapping of H3K4me1+/H3K4me3- peaks shows that putative bystanders for CAMPs contain ~4 times more active enhancers than for non-conserved pairs (p=0.005, KS test, Figure 4A), and with a higher density (p=0.003, Figure 4B). These results suggest that conserved GRBs have complex *cis*-regulatory landscapes in zebrafish development. Importantly, we found that H3K27me3 is significantly increased in *trans-dev* but not in bystander genes in whole zebrafish embryos (Figure 4C-E). This suggests that, globally, the *trans-dev* genes of GRBs, but not the enhancer-containing bystanders, have complex tissue-specific expression patterns (Akkers et al. 2009).

A striking case of conserved GRB involves the ISL LIM homeobox (*Isl*). *Isl* plays important conserved roles in animal development, in particular in neuron ontogeny in diverse phyla (Thor, Thomas 1997; Jackman, Langeland, Kimmel 2000; Voutev et al. 2009; Liang et al. 2011). We found that *Isl* genes are linked to *Scaper*/*ssp3* (S-phase cyclin A-associated protein in the ER) genes in nearly all studied species, from sponges to humans, for over 1,000 MY of evolution (Figure 5A). In human, *SCAPER* spans ~500 Kbp and contains 16 introns longer than 10 Kbp, as expected for a bystander. Furthermore, the expression patterns of these two genes are very different in flies (r=0.08), humans (r=-0.07) and zebrafish (Figure 5B). In 24hpf zebrafish embryos, *isl2a* is expressed in specific cephalic neuronal domains and in discrete neurons in the spinal cord (arrow), and in hindgut (red arrow) – a pattern conserved in other animal lineages (Jackman, Langeland, Kimmel 2000; Gelbart, Emmert 2010; Graveley et al. 2011) – whereas *scaper* is not expressed at this stage.

In zebrafish, ChIP-seq data for H3K4me1 showed three potential enhancers within *scaper* at 24hpf (red bars in Figure 5C), whereas H3K4me3 indicated that *isl2a*'s promoter is active but *scaper*'s is not (Figure 5C), suggesting that the active enhancers within *scaper* may be acting on *isl2a*. We thus used the GFP reporter vector ZED (Bessa et al. 2009) to examine the enhancer activity of the three H3K4me1+/H3K4me3- peaks in stable zebrafish transgenic lines. One of them (red asterisk in Figure 5C) promoted GFP expression in different neuron populations and hindgut, two domains co-expressing the endogenous *Isl2a* protein, as shown by ISH and immunocolocalization (Figure 5B and Figure S3). These results strongly suggest that at least this *scaper*-contained enhancer is acting specifically on *isl2a*, and provide a plausible explanation for the conservation of this gene association.

Another interesting case is the homeobox transcription factor *Otx* (Figure 6). This gene is involved in establishing a deeply conserved early anterior-posterior (A-P) brain patterning (Simeone et al. 1992; Hirth et al. 2003; Castro et al. 2006; Irimia et al. 2010), and is expressed in the anterior nervous system of nearly all studied bilaterians (Simeone et al. 1992; Williams, Holland 1996; Hirth et al. 2003; Lowe et al. 2003; Scholpp et al. 2007). We found that *Otx* is linked to *Ehbp1* (EH domain binding protein 1) in nearly all species, from placozoans to humans, spanning over 700 MY (Figure 6A). As in the case of *Isl-Scaper*, the expression of *OTX1* and *EHBP1* is very different in human (r=-0.12) and in zebrafish (Figure 6B). During early zebrafish development, only *otx1b* is expressed in the embryo, mainly in the cephalic region. In addition, zebrafish *ehbp1* also has several long introns that contain potential enhancers, as indicated by high levels of H3K4me1 without H3K4me3 (Figure 6C). We therefore generated stable transgenic zebrafish reporter lines for three of these H3K4me1 peaks, one of which drove strong and consistent expression to the classical conserved *Otx* anterior domain in the central nervous system. Next, to test if the remaining *ehbp1* intronic sequences may also contain other *otx1b cis*-regulatory elements, we performed chromosome conformation capture (3C) assays (Hagège et al. 2007) to explore physical interaction

between several regions of the *ehbp1* intron (red arrows in Figure 6C) and the *otx1b* promoter. This assay confirmed the interaction of the identified enhancer, and revealed another potential *cis*-regulatory element within the *ehbp1* intron that clearly contacted the *otx1b* promoter (Figure 6C). Collectively, these results show that *ehbp1* contains *cis*-regulatory elements that specifically regulate the expression of *otx1b* in deeply conserved domains.

Discussion

We have reported a large number of genes belonging to Conserved Ancestral Microsyntenic Pairs (CAMPs) shared across several deeply diverged metazoan lineages. In contrast to the few previously described cases of deeply conserved microsynteny, nearly all of which involve tandemly duplicated gene pairs or clusters, we found nearly 600 pairs of unrelated genes that are closely physically linked in several major bilaterian lineages spanning over 600 MY of animal evolution (half of them already present in the non-bilaterian ancestors, 700-1,000 MY ago). Moreover, these numbers are likely underestimates, since most genome sequence assemblies used in our study are highly fragmented into relatively small scaffolds (Table S1). Thus, much of the microsynteny in most organisms could not be truly evaluated and was conservatively considered to be non-conserved.

These results are quite unexpected in light of previous studies of pairwise conservation. Current and previous comparisons between pairs of distantly related species show very little conservation of local microsynteny (e.g. only ~1-5% between humans and other lineages, Figure S4). Indeed, even studies of slow-evolving metazoans, which found conservation of chromosome-level linkage (e.g., genes on the same chromosome in humans also tend to be on the same chromosome in cnidarians), found very little or no microsyntenic conservation across several species (Putnam et al. 2007; Putnam et al. 2008; Srivastava et al. 2008; Srivastava et al. 2010). Our finding that comparable fractions of gene linkages (e.g., including 1-2% of genes in humans, Table

S1) have been independently conserved in several different lineages is therefore quite unexpected, and implies that the same specific subset of gene linkages have been actively retained in widely diverged lineages (including examples even in typically fast-evolving species (Chavali et al. 2011; Lv, Havlak, Putnam 2011; Weber, Hurst 2011)). Accordingly, we have provided extensive evidence that many of these associations have been conserved due to functional reasons. First, many CAMPs show high coexpression levels in humans and/or flies, suggesting that they may share *cis*-regulatory inputs resulting in coordinated transcription. Second, we found more than one hundred putative ancient GRBs, often involving important developmental regulators, substantially adding to the three previously known trans-phyletic GRB (Wang et al. 2007; Irimia et al. 2012; Maeso et al. 2012). The deep conservation of GRBs suggests that they may underlie the remarkable conservation of some developmental genetic programs, such as the A-P patterning of the bilaterian central nervous system (in the case of *Otx-Ehbp1*) or neuronal ontogeny (*Isl-Scaper*). At the same time, the conservation of some GRBs in lineages with very different body plans or cell types is intriguing. For example, in the case of *Isl*, it is not clear which common regulatory role may be responsible for the conservation of the association between bilaterians and sponges, which have no proper neurons (Hooper, Van Soest 2002; Sakarya et al. 2007).

The phylogenetic distribution of CAMPs also shows that, despite their deep conservation, many of the associations have been repeatedly lost in different lineages. This was observed even for extremely conserved CAMPs such as the coexpressed mitochondrial chaperonins *HSPE1-HSPD1* (lost in flies), and the GRBs described in detail in the present study (Figures 5A and 6A). Different causes may underlie the loss of these associations. For example, major modification of body plans may render some of the regulatory constraints obsolete, and thus the microsynteny can be lost presumably without selective cost (as in the case of *Hox* clusters (Duboule 2007)). Alternatively, associations within GRBs could be disentangled through the acquisition of genetic redundancy (McEwen et al. 2006; Kikuta et al. 2007; Navratilova et al. 2010; Goode et al. 2011; Maeso et al.

2012). Complex genes are often regulated by a redundant set of *cis*-regulatory elements, which increase robustness, and are continuously evolving (Jeong et al. 2006; Hong, Hendrix, Levine 2008; Frankel et al. 2010; Perry et al. 2010; Schmidt et al. 2010). For example, if a new, redundant *cis*-regulatory module arises outside the bystander gene (e.g. in the intergenic region), the bystander gene could be now translocated without affecting the *trans-dev* gene's regulation. In a more complex scenario, duplication of the gene pair (either by whole genome or segmental duplication) can aid this process. In coregulated gene pairs, each gene of the pair could be reciprocally lost from one of the duplicated regions while keeping the common *cis*-regulatory elements in both, therefore becoming two non-associated genes with fully functional regulatory landscapes. In the case of GRBs, the coding sequences of extra bystander gene copies may be erased, while the *trans-dev*-associated regulatory elements are conserved (McEwen et al. 2006; Kikuta et al. 2007; Maeso et al. 2012).

Finally, the extent of ancient conserved microsynteny we have uncovered is even more striking when considering that the ancestral bilaterian likely had fewer than 10,000 unique genes (Miller, Ball 2009): thus, over 12% of these are involved in close microsyntenic relationships conserved in multiple lineages to this day. It can thus be argued that microsynteny is among the most conserved features of metazoan genomes, and that ancient *cis*-regulatory inputs may be far more common than currently appreciated (Royo et al. 2011). We expect that as metazoan genomes continue to be sequenced at an ever-faster rate, many more microsyntenic relationships will be discovered, and many more details surrounding their role as key components of the genome's architecture will be revealed.

Methods

Genome-wide search for deeply conserved gene pairwise associations

We downloaded full genome annotations for the initial 13 species (File S1). For multiple-transcript genes we used only the longest protein isoform. Homologs were identified by pairwise blastP (e-value < 10^{-5} without filtering for low complexity sequences (-F F)).

For each pair of immediately adjacent protein-coding genes (excluding non-coding RNA genes, pseudogenes, etc) in each genome, we asked whether the first or second best BLAST hits for both proteins were also neighboring in each of the 12 other genomes. This was defined as on the same chromosome/scaffold with ≤4 intervening genes, a threshold chosen to deal with: (i) common annotation errors (e.g., spurious automatic gene models, split genes), and (ii) the fact that GRBs often include non-adjoining genes (Kikuta et al. 2007). Cutoffs of 3-10 intervening genes yielded similar results, Figure S5). Based on randomization simulations (see below), we considered a gene pair conserved if it was linked in four total lineages (except vertebrates, each species was considered a distinct lineage, since pairwise species comparisons showed no enrichment for pair conservation due to phylogenetic proximity, except within vertebrates, Figure S4). To determine 'conservation' by chance, we randomized gene order within each chromosome/scaffold for each of the 12 species, with 100 replicates. These simulations indicated that a cutoff of ≥ 4 lineages with ≤ 4 intervening genes yields a false positive discovery rate of <0.0002 per gene pair for all species (Table S1).

Generation of a unique dataset of ancient microsyntenic gene pairs

Conserved pairs for each of the 13 species were then merged into a single dataset of non-redundant groups. Each unique group contained the syntenic pairs for all species in which it was conserved, including duplicates of the pairs within species (e.g. paralogons in vertebrates resulted from the two rounds of whole genome duplication), if any. Next, we assessed whether the genes were related (paralogous pairs), and filters were applied to account for reciprocal blast consistency and common annotation errors (see File S1 for details).

To reconstruct the history of linked duplicate genes, we performed Bayesian phylogenetic inferences for genes for all species for each pair, to distinguish (i) independent tandem duplications (clustering of genes by species on phylogenetic trees) and (ii) retained ancestral linkage. Statistical significance was tested by comparison to randomized trees with the same topology (File S1 for details).

We also studied disruption of CAMPs. First, we assessed conservation in four additional species: *S. kowalevskii*, *O. dioica*, *A. queenslandica* and *C. owczarzaki* using tBLASTN against the assembled contigs (File S1). Then, we applied parsimony to infer the number of CAMPs that were present at each node of a consensus phylogenetic tree, and estimated the fraction that were disrupted along each branch.

Coexpression of CAMPs using microarray data

We downloaded the full set of experiments from Affymetrix Human Genome U133 Plus 2.0 Array and Affymetrix Drosophila Genome 2.0 Array (GEO accession numbers GPL570 and GPL1322). After excluding experiments with missing probes, we obtained a matrix with 54,608 probes with data for 23,941 experiments in humans, and 13,935 probes with data for 1,909 experiments in Drosophila. The expression levels were converted to ranks within each experiment to correct for different measurement and normalization methods.

Of the CAMPs with no intervening genes, 279 in humans and 27 in Drosophila had at least one probe for each gene. For these, we calculated the correlation coefficient for gene expression between the two genes. For comparisons, we generated two control sets for each conserved pair: (i) 100 non-paralogous, non-conserved gene pairs with the same orientation and the most similar intergenic distances; (ii) 100 pairs of two randomly selected non-syntenic genes. Then, the correlation coefficient of each conserved pair was ranked with respect to each of its two control sets. When multiple probes covered a gene, we used the combination of probes that gave the highest

correlation (for both test and control sets). Using the average between probes gave the same pattern of higher co-expression between conserved pairs.

Study of genome structure

Orientation and intergenic distance were calculated from GFF/GTF annotation files, using the stable transcript for Ensembl genomes, and the best gene models for other genomes. We merged all gene isoforms into a single intron-exon structure to determine intron number/lengths. Global sequence conservation was calculated using the phastCons46wayPlacental scores from UCSC Comparative track (http://genome.ucsc.edu/). Only sites with scores >0 (confidently aligned across genomes) were used for calculations. HCNRs were obtained from two sources: (i) *vistaEnhancers* track at UCSC, and (ii) ancient conserved non-coding elements (aCNEs, (Lee et al. 2011)). Active strong enhancers and insulators for nine human cell lines were obtained from (Ernst et al. 2011). Data for all cell lines were merged into a unique, non-redundant set of coordinates.

To study GRBs, we defined developmental (*trans-dev*) genes as genes with GO terms embryo development (GO:0009790) and/or organ development (GO:0048513). For comparison, we also defined two sets of CAMPs composed of two non-developmental genes, based on the analyses of coexpression presented above: pairs of highly coexpressed non-developmental genes ($r>0.40$, likely enriched in gene pairs conserved due to co-regulatory reasons) and of weakly coexpressed non-developmental genes ($r<0.05$). In addition, each of these conserved pair sets had its corresponding control set, consisting of similar non-paralogous syntenic pairs (i.e. bystander plus *trans-dev* or highly/lowly coexpressed) that are not evolutionarily conserved.

Experimental analyses in zebrafish embryos

Chromatin immunoprecipitation (ChIP) was performed following the protocol described in (Wardle et al. 2006) with minor modifications, and Genome Analyzer (Illumina) ChIP-seq was performed as

described in (Bogdanović et al. 2012) (see File S1 for details). Highly enriched regions (peaks) of histone methylation were obtained by the MACS (v.1.3.3) algorithm (Zhang et al. 2008) using standard settings with one modification (mfold = 20). Twenty-five randomly selected peaks were verified by qPCR and compared to their random controls (false positive discovery rate (FDR) < 0.04). The PCRs were performed on 1:50 dilutions of the ChIP samples using the C1000 Thermal Cycler (BioRad).

For each H3K4me1-positive peak that was tested in stable zebrafish transgenic assays, we designed primers to span the whole region plus ~100nt at each side (primer sequences are available in Table S5). Specific details for cloning, preparation and injection of candidate enhancer sequences into zebrafish eggs, as well as for *in situ* hybridization and immunostaining, are provided in File S1; in all cases, previously described protocols were followed with minor modifications (Kawakami 2004; Tena et al. 2007; Bessa et al. 2009). 3C assays were performed as referred in ((Hagège et al. 2007; Tena et al. 2011), see File S1). A set of locus-specific primers (Table S5) was designed to perform qPCRs to measure relative enrichment in each ligation product. Negative control primers were designed ~30 Kbp upstream and downstream the regions of interest. PCR values were normalized with primers for *Ercc3*.

Data Access

Short read data has been submitted to GEO (H3K27me3, GSE35050, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=lrazdiegwqoqyvi&acc=GSE35050; H3K4me1 and H3K4me3, GSE32483, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=vnyhbkqayaioczc&acc=GSE32483).

Acknowledgements

MI, MSA, SWR and HBF were funded by NIH grant 1R21HG005240-01A1. HBF is a Alfred P. Sloan Fellow and Pew Scholar in the Biomedical Sciences. JJT, AF-M, OB, EC-M, and OJLG-S were funded by grants BFU2010-14839, CSD2007-00008 and Proyecto de Excelencia CVI-3488.

References

Adachi, N, MR Lieber. 2002. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**:807-809.

Akkers, RC, SJ van Heeringen, UG Jacobi, EM Janssen-Megens, KJ Françoijs, HG Stunnenberg, GJ Veenstra. 2009. A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Dev Cell* **17**:425-434.

Baldwin, W, P Marko, D Nelson. 2009. The cytochrome P450 (CYP) gene superfamily in Daphnia pulex. *BMC Genomics* **10**:169.

Becker, TS, B Lenhard. 2007. The random versus fragile breakage models of chromosome evolution: a matter of resolution. *Mol Genet Genomics* **278**:487-491.

Bessa, J, JJ Tena, E de la Calle-Mustienes, A Fernández-Miñán, S Naranjo, A Fernández, L Montoliu, A Akalin, B Lenhard, F Casares, *et al.* 2009. Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev Dyn* **238**:2409-2417.

Bhutkar, A, SW Schaeffer, SM Russo, M Xu, TF Smith, WM Gelbart. 2008. Chromosomal Rearrangement Inferred From Comparisons of 12 Drosophila Genomes. *Genetics* **179**:1657-1680.

Bogdanović, O, A Fernandez-Miñán, JJ Tena, E de la Calle-Mustienes, C Hidalgo, I van Kruysbergen, SJ van Heeringen, GJ Veenstra, JL Gómez-Skarmeta. 2012. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res*:In press.

Bourque, G, EM Zdobnov, P Bork, PA Pevzner, G Tesler. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res* **15**:98-110.

Castro, LFC, SLK Rasmussen, PWH Holland, ND Holland, LZ Holland. 2006. A Gbx homeobox gene in amphioxus: Insights into ancestry of the ANTP class and evolution of the midbrain/hindbrain boundary. *Dev Biol* **295**:40-51.

Chavali, S, DA Morais, J Gough, MM Babu. 2011. Evolution of eukaryotic genome architecture: Insights from the study of a rapidly evolving metazoan, Oikopleura dioica: Non-adaptive forces such as elevated mutation rates may influence the evolution of genome architecture. *Bioessays* **33**:592-601.

Davila-Lopez, M, JJ Martinez-Guerra, T Samuelsson. 2010. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS One* **5**:e10654.

Denoeud, F, S Henriet, S Mungpakdee, JM Aury, C Da Silva, H Brinkmann, J Mikhaleva, LC Olsen, C Jubin, C Cañestro, *et al.* 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**:1381-1385.

Doolittle, WF. 1978. Genes in pieces: were they ever together? *Nature* **272**:581-582.

Duboule, D. 2007. The rise and fall of Hox gene clusters. *Development* **134**:2549-2560.

Engstrom, PG, SJ Ho Sui, O Drivenes, TS Becker, B Lenhard. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res*. **17**:1898-1908.

Ernst, J, P Kheradpour, TS Mikkelsen, N Shoresh, LD Ward, CB Epstein, X Zhang, L Wang, R Issner, M Coyne, *et al.* 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**:43-49.

Fischer, G, EP Rocha, F Brunet, M Vergassola, B Dujon. 2006. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet* **2**:e32.

Frankel, N, GK Davis, D Vargas, S Wang, F Payre, DL Stern. 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**:490-493.

Fukuoka, Y, H Inaoka, IS Kohane. 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. *BMC Genomics* **5**:4.

Gelbart, WM, DB Emmert. 2010. FlyBase High Throughput Expression Pattern Data Beta Version.

Goode, DK, HA Callaway, GA Cerda, KE Lewis, G Elgar. 2011. Minor change, major difference: divergent functions of highly conserved cis-regulatory elements subsequent to whole genome duplication events. *Development* **138**:879-884.

Graveley, BR, AN Brooks, JW Carlson, MO Duff, JM Landolin, L Yang, CG Artieri, Ba van, M.J, N Boley, BW Booth, *et al.* 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**:473-479.

Hagège, H, P Klous, C Braem, E Splinter, J Dekker, G Cathala, W de Laat, T Forné. 2007. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc* **2**:1722-1733.

Hansen, JJ, P Bross, M Westergaard, MN Nielsen, H Eiberg, AD Børglum, J Mogensen, K Kristiansen, L Bolund, N Gregersen. 2003. Genomic structure of the human mitochondrial chaperonin genes: HSP60 and HSP10 are localised head to head on chromosome 2 separated by a bidirectional promoter. *Hum Genet* **112**:71-77.

Hirth, F, L Kammermeier, E Frei, U Walldorf, M Noll, H Reichert. 2003. An urbilaterian origin of the tripartite brain: developmental genetic insights from Drosophila. *Development* **130**:2365-2373.

Hong, JW, DA Hendrix, MS Levine. 2008. Shadow enhancers as a source of evolutionary novelty. *Science* **321**:1314.

Hooper, JNA, RWM Van Soest. 2002. Systema Porifera: A Guide to the classification of sponges. New York: Kluwer Academic/Plenum Publishers.

Hurst, LD, C Pal, MJ Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* **5**:299-310.

Hurst, LD, EJ Williams, C Pal. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet* **18**:604-606.

Irimia, M, I Maeso, D Burguera, M Hidalgo-Sánchez, L Puelles, J Garcia-Fernàndez, SW Roy, JL Ferran. 2011. Contrasting 5' and 3' Evolutionary Histories and Frequent Evolutionary Convergence in Meis/hth Gene Structures. *Genome Biol Evol* **3**:551-564.

Irimia, M, I Maeso, J Garcia-Fernandez. 2008. Convergent evolution of clustering of Iroquois homeobox genes across metazoans. *Mol Biol Evol* **25**:1521-1525.

Irimia, M, C Piñeiro, I Maeso, JL Gómez-Skarmeta, F Casares, J Garcia-Fernàndez. 2010. Conserved developmental expression of Fezf in chordates and Drosophila and the origin of the Zona Limitans Intrathalamica (ZLI) brain organizer. *EvoDevo* **1**:7.

Irimia, M, SW Roy. 2008. Spliceosomal introns as tools for genomic and evolutionary analysis. *Nucleic Acids Res.* **36**:1703-1712.

Irimia, M, JL Royo, D Burguera, I Maeso, JL Gómez-Skarmeta, J Garcia-Fernandez. 2012. Comparative genomics of the Hedgehog loci in chordates and the origins of Shh regulatory novelties. *Sci Rep* **2**:433.

Jackman, WR, JA Langeland, CB Kimmel. 2000. islet reveals segmentation in the Amphioxus hindbrain homolog. *Dev Biol* **220**:16-26.

Jeong, Y, K El-Jaick, E Roessler, M Muenke, DJ Epstein. 2006. A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* **133**:761-772.

Kawakami, K. 2004. Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element. *Methods Cell Biol* **77**.

Kent, WJ, AM Zahler. 2000. Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. *Genome Res.* **10**:1115-1125.

Kikuta, H, M Laplante, P Navratilova, AZ Komisarczuk, PG Engstrom, D Fredman, A Akalin, M Caccamo, I Sealy, K Howe, *et al.* 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**:545-555.

Koonin, EV, YI Wolf. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet* **11**:487-498.

Lee, AP, SY Kerk, YY Tan, S Brenner, B Venkatesh. 2011. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol Biol Evol* **28**:1205-1215.

Lee, JM, EL Sonnhammer. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13**:875-882.

Liang, X, MR Song, Z Xu, GM Lanuza, Y Liu, T Zhuang, Y Chen, SL Pfaff, SM Evans, Y Sun. 2011. Isl1 is required for multiple aspects of motor neuron development. *Mol Cell Neurosci* **47**:215-222.

Lowe, CJ, M Wu, A Salic, L Evans, E Lander, N Stange-Thomann, CE Gruber, J Gerhart, M Kirschner. 2003. Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* **113**:853-865.

Lv, J, P Havlak, NH Putnam. 2011. Constraints on genes shape long-term conservation of macro-synteny in metazoan genomes. *BMC Bioinformatics* **12**:S11.

Lynch, M. 2007. The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates.

Lynch, M, JS Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:1151-1155.

Maeso, I, M Irimia, JJ Tena, E González-Pérez, D Tran, V Ravi, B Venkatesh, S Campuzano, JL Gómez-Skarmeta, J Garcia-Fernàndez. 2012. An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome Res* **22**:642-655.

Maeso, I, SW Roy, M Irimia. 2012. Widespread recurrent evolution of genomic features. *Genome Biol Evol* **4**:486-500.

Margueron, R, D Reinberg. 2010. Chromatin structure and the inheritance of epigenetic information. *Nat Rev Genet* **11**:285-296.

McEwen, GK, A Woolfe, D Goode, T Vavouri, H Callaway, G Elgar. 2006. Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Res* **16**:451-465.

Michalak, P. 2008. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* **91**:243-248.

Miller, DJ, EE Ball. 2009. The gene complement of the ancestral bilaterian - was Urbilateria a monster? *J Biol* **8**:89.

Navratilova, P, D Fredman, B Lenhard, TS Becker. 2010. Regulatory divergence of the duplicated chromosomal loci sox11a/b by subpartitioning and sequence evolution of enhancers in zebrafish. *Mol Genet Genomics* **283**:171-184.

Oliver, B, T Misteli. 2005. A non-random walk through the genome. *Genome Biol* **6**:214.

Perry, MW, AN Boettiger, JP Bothma, M Levine. 2010. Shadow enhancers foster robustness of Drosophila gastrulation. *Curr Biol* **20**:1562-1567.

Poyatos, JF, LD Hurst. 2007. The determinants of gene order conservation in yeasts. *Genome Biol* **8**:R233.

Putnam, N, T Butts, DEK Ferrier, RF Furlong, U Hellsten, T Kawashima, M Robinson-Rechavi, E Shoguchi, A Terry, JK Yu, *et al.* 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**:1064-1071.

Putnam, NH, M Srivastava, U Hellsten, B Dirks, J Chapman, A Salamov, A Terry, H Shapiro, E Lindquist, VV Kapitonov, *et al.* 2007. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science* **317**:86-94.

Quijano, C, P Tomancak, J Lopez-Marti, M Suyama, P Bork, M Milan, D Torrents, M Manzanares. 2008. Selective maintenance of Drosophila tandemly arranged duplicated genes during evolution. *Genome Biol* **9**:R176.

Royo, JL, I Maeso, M Irimia, F Gao, IS Peter, CS Lopes, S D'Aniello, F Casares, EH Davidson, J Garcia-Fernández, *et al.* 2011. Transphyletic conservation of developmental regulatory state in animal evolution. *Proc Natl Acad Sci USA* **108**:14186-14191.

Sakarya, O, KA Armstrong, M Adamska, M Adamski, IF Wang, B Tidor, BM Degnan, TH Oakley, KS Kosik. 2007. A post-synaptic scaffold at the origin of the animal kingdom. *PLoS One* **6**:e506.

Schmidt, D, MD Wilson, B Ballester, PC Schwalie, GD Brown, A Marshall, C Kutter, S Watt, CP Martinez-Jimenez, S Mackay, *et al.* 2010. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* **328**:1036-1040.

Scholpp, S, I Foucher, N Staudt, D Peukert, A Lumsden, C Houart. 2007. Otx1l, Otx2 and Irx1b establish and position the ZLI in the diencephalon. *Development* **134**:3167-3176.

Simeone, A, D Acampora, M Gulisano, A Stornaiuolo, E Boncinelli. 1992. Nested expression domains of four homeobox genes in developing rostral brain. *Nature* **358**:687-690.

Srivastava, M, E Begovic, J Chapman, NH Putnam, U Hellsten, T Kawashima, A Kuo, T Mitros, A Salamov, ML Carpenter, *et al.* 2008. The Trichoplax genome and the nature of placozoans. *Nature* **454**:955-960.

Srivastava, M, O Simakov, J Chapman, B Fahey, MEA Gauthier, T Mitros, GS Richards, C Conaco, M Dacre, U Hellsten, *et al.* 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**:720-726.

Stein, LD, Z Bao, D Blasiar, T Blumenthal, MR Brent, N Chen, A Chinwalla, L Clarke, C Clee, A Coghlan, *et al.* 2003. The genome sequence of Caenorhabditis briggsae: A platform for comparative genomics. *PLoS Biol* **1**:E45.

Tena, JJ, ME Alonso, E de la Calle-Mustienes, E Splinter, W de Laat, M Manzanares, JL Gómez-Skarmeta. 2011. An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and co-regulation. *Nat Commun* **In press**.

Tena, JJ, A Neto, E de la Calle-Mustienes, C Bras-Pereira, F Casares, JL Gomez-Skarmeta. 2007. Odd-skipped genes encode repressors that control kidney development. *Dev Biol* **301**:518-531.

Thomas, JH. 2007. Rapid Birth-Death Evolution Specific to Xenobiotic Cytochrome P450 Genes in Vertebrates. *PLoS Genet* **3**:e67.

Thor, S, JB Thomas. 1997. The Drosophila islet gene governs axon pathfinding and neurotransmitter identity. *Neuron* **18**:397-409.

Turner, BM. 2007. Defining an epigenetic code. *Nat Cell Biol* **9**:2-6.

Visel, A, S Minovitsky, I Dubchak, LA Pennacchio. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucl Acids Res* **35**:D88-92.

Voutev, R, R Keating, EJ Hubbard, LG Vallier. 2009. Characterization of the Caenorhabditis elegans Islet LIM-homeodomain ortholog, lim-7. *FEBS Lett* **583**:456-464.

Wang, W, J Zhong, B Su, Y Zhou, YQ Wang. 2007. Comparison of Pax1/9 Locus Reveals 500-Myr-Old Syntenic Block and Evolutionary Conserved Noncoding Regions *Mol Biol Evol* **24**:784-791.

Wardle, FC, DT Odom, GW Bell, B Yuan, TW Danford, EL Wiellette, E Herbolsheimer, HL Sive, RA Young, JC Smith. 2006. Zebrafish promoter microarrays identify actively transcribed embryonic genes. *Genome Biol* **7**:R71.

Weber, CC, LD Hurst. 2011. Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation. *Genome Biol* **12**:R23.

Williams, NA, PW Holland. 1996. Old head on young shoulders. *Nature* **383**:490.

Woolfe, A, M Goodson, DK Goode, P Snell, GK McEwen, T Vavouri, SF Smith, P North, H Callaway, K Kelly, *et al.* 2005. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biology* **3**:e7.

Zhang, Y, T Liu, CA Meyer, J Eeckhoute, DS Johnson, BE Bernstein, C Nusbaum, RM Myers, M Brown, W Li, *et al.* 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**:R137.

Figure captions

Figure 1 – Functional causes of conserved microsynteny and phylogenetic distribution of conserved gene pairs

A) Co-regulation: two neighbor genes share one or more regulatory elements. Chromosomal breakpoints in the intergenic region of the two genes are opposed by selection since it would affect the coordinated expression of the two genes. B) Genomic regulatory block (GRB): a bystander gene (green) contains regulatory elements in its introns that target a neighbor gene (red), often a *trans-dev* gene with key regulatory roles in animal development. The breakage of the association would result in affected regulation of the *trans-dev* gene. Modified from (Becker, Lenhard 2007). C) Consensus phylogeny of the studied metazoans, showing the total number of gene pairs in each species (in parenthesis), and the minimum number of pairs at each node inferred by parsimony (in boxes). Branch lengths correspond to the fraction of gene pairs lost out of the total pairs present in the last common ancestor. Black dots at basal nodes indicate that branch length could not be estimated.

Figure 2 – Predictive features of co-regulated gene pairs

A) Percentage of gene pairs oriented in 3'-3' (convergently transcribed), 5'-3' or 3'-5' (co-directionally transcribed) or 5'-5' (divergently transcribed) among human conserved gene pairs (dark grey) and control set (all non-conserved, non-paralogous gene pairs; light grey). Conserved pairs are highly enriched in 5'-5' orientation ($p=1x10^{-9}$, χ^2 test). B) Percentage of pairs with fully conserved orientation across species for human conserved gene pairs; 5'-5' pairs are more conserved. C) Cumulative plot of ranks of ranked correlation coefficients of conserved pairs among the control dataset (non-conserved neighbor pairs with the same orientation and similar intergenic distance; black) and random gene pairs (grey). The red line shows the expected cumulative pattern if co-expression was similar to the controls, and the area above the line the excess of highly co-

expressed pairs. D) Percentage of conserved pairs with ranks in each of the decimal bins. The red line (10%) indicates the expected values. Conserved pairs were significantly enriched in ranks of the last decimal bin (91-100, red arrow head).

Figure 3 – Different functional and evolutionary features of intragenic sequences of conserved pairs

A) Average intron number for each gene type. Dark grey: conserved pairs; light grey: control set. Putative human bystander genes (non-developmental genes associated with *trans-dev* genes in conserved pairs) have significantly more introns in average than the control set ($p=7x10^{-4}$, KS test), whereas associated *trans-dev* genes have fewer ($p=6x10^{-4}$, KS test). Genes in highly co-expressed or lowly co-expressed conserved pairs have similar average intron numbers than the control set. B) Average intron length (in Kbp). Putative bystander genes have significantly longer introns than the other genes ($p=2x10^{-6}$, KS test). C) Average number of introns longer than 10 Kbp. Bystander genes have nearly three times more long introns than other genes ($p=2x10^{-8}$). D) Average mammalian-wide phastCons score of intronic regions of different types of conserved gene pairs. Both *trans-dev* and bystander genes from conserved pairs show higher intronic sequence conservation than the non-conserved genes ($p=0.002$ and $p=3x10^{-5}$, KS test), whereas conserved highly co-expressed non-developmental gene pairs show higher conservation than the control set at the intergenic region. E) Density of ancient conserved non-coding elements (aCNEs) per Mbp in introns of different types of conserved gene pairs. F) Density of functionally defined strong enhancers per Kbp in introns of different types of conserved gene pairs. Error bars correspond to standard errors.

Figure 4 – Differential epigenetic marks in in bystander and *trans-dev* genes in early zebrafish development.

A, B) Average number (A) and density (peaks/Kbp, B) of H3K4me1+/H3K4me3- peaks (putative active enhancers) in 24hpf zebrafish embryos within *trans-dev*, bystander and other non-developmental genes for conserved pairs (dark grey) and for non-conserved pairs (light gray). Error bars correspond to standard errors. C) Average number of reads for H3K27me3 ChIP-seq around transcription start sites show an specific deposition of this epigenetic mark in *trans-dev* genes. D, E) Two examples of differential distribution of H3K27me3 in a *trans-dev* gene (*foxc1b* (D)) and *fgf20a* (E)) and its non-developmental partner (*gmds* (D) and *efha2* (E)).

Figure 5 – Functional characterization of the GRB of *isl2a-scaper* in zebrafish

A) Phylogenetic distribution of the GRB across the studied metazoan species. The GRB was only not conserved in *C. elegans* and *N. vectensis*. B) Upper and middle panels are 24 hours post fertilization (hpf) zebrafish embryos showing the expression patterns of *isl2a* and *scaper*. *isl2a* is expressed in discrete neurons in the spinal cord (arrow) and in the proctodeum (arrowhead) while *scaper* is not detectable at this stage. The third panel is an embryo at a similar developmental stage showing GFP expression promoted by an enhancer located within the intron of *scaper* (asterisk in C). This enhancer is active in *isl2a*-expressing territories. The bottom panel shows immunodetection of the transgenic GFP (green), the endogenous Islet proteins (red) and the overlay between the two showing coexpression. C) Distribution of H3K4me3, H3K4me1 and H3K27me3 tracks along the *isl2a-scaper* GRB in 24 hpf zebrafish embryos. H3K4me1 peaks tested for enhancer activity in zebrafish stable transgenic lines are shaded in red. H3K4me3 and H3K27me3 distribution show that *scaper* is inactive and this stage and that *isl2a* is tissue specific, suggesting that the H3K4me1-positive enhancer may be acting on the active *isl2a* promoter. Below, conservation track from UCSC browser.

Figure 6 – Functional characterization of the GRB of *otx1b-ehbp1* in zebrafish

A) Phylogenetic distribution of the GRB across the studied metazoan species. The GRB was not conserved in *D. melanogaster*, *C. elegans* and *N. vectensis*. B) Upper and middle panels are zebrafish embryos at 24 hpf showing the expression of *otx1b* and *ehbp1* genes. *otx1b* is detected in the anterior brain while *ehbp1* is expressed in the yolk. Lower panel show GFP expression promoted by an enhancer located within the intron of *ehbp1* (asterisk in C). This enhancer is active in most tissues expressing *otx1b*. C) Distribution of H3K4me3, H3K4me1 and H3K27me3 tracks along the *otx1b-ehbp1* GRB in 24 hpf zebrafish embryos. H3K4me1 peaks tested for enhancer activity in zebrafish stable transgenic lines are shaded in red. The regions that physically interact in 3C assays with the *otx1b* promoter are shaded in blue. H3K27me3 distribution indicates that *otx1b*, but not *ehbp1*, has tissue specific expression, hence H3K4me1 enhancers located in *ehbp1* introns are likely acting on *otx1b*. Below, conservation track from UCSC browser. D) Graph showing a 3C experiment to detect interaction between different *ehbp1* intronic regions and the *otx1b* promoter in 24 hpf embryos. A fixed primer (yellow arrow) was set at the *otx1b* promoter and 7 regions were assayed for interaction with that promoter using different primers (red arrows) distributed along the *ehbp1* intronic genomic area. The highest crosslinking frequency value is set to 1. Error bars indicate SE (n=3).

Supplementary Files

File S1 – Supplementary Methods

Figure S1 – Example of convergently evolved paralogous pairs

A) Example of a group of genes likely resulted by independent tandem gene duplication (*c16orf5*). Abbreviations: Bfl, *Branchiostoma floridae*; Lgi, *Lottia gigantea*; Cte, *Capitella teleta*; Dme, *Drosophila melagonaster*; Tad, *Trichoplax adhaerens*; Cel, *Caenorhabditis elegans*.

Figure S2 – Coexpression analysis for 27 fly gene pairs

A) Cumulative plot of ranks of ranked correlation coefficients of conserved pairs among the control dataset (non-conserved neighbor pairs with the same orientation and similar intergenic distance). The red line shows the expected cumulative pattern if co-expression was similar to the controls, and the area above the line the excess of highly co-expressed pairs. D) Percentage of conserved pairs with ranks in each of the decimal bins. The red line (10%) indicates the expected values. Conserved pairs were significantly enriched in ranks of the last decimal bin (91-100).

Figure S3 – Coexpression of endogenous ISL and GFP driven by a *scaper2a*-contained enhancer

A) Expression of GFP (green), Islet (red) and coexpression (merge, right panel) in motoneurons (arrows) and sensory Rohon-Beard cells (arrow heads). B) Expression of GFP (green), Islet (red) and coexpression (merge, right panel) in neurons in the basal ganglia (arrows).

Figure S4 – Pairwise conservation of syntenic pairs between human and each of the studied species

Histogram of the percentage of conserved pairs between human and the other 12 species, and also mouse, opossum and fugu, suggesting that rearrangements seem to have reached saturation in pairwise comparisons with invertebrate comparisons, but not within vertebrates (and were thus considered as a single lineage in this study). Mmu, *Mus musculus*; Mdo, *Monodelphis domestica*; Gga, *Gallus gallus*; Xtr, *Xenopus tropicalis*; Tru, *Takifugu rubripes*; Cin, *Ciona intestinalis*; Bfl, *Branchiostoma floridae*; Spu, *Strongylocentrotus purpuratus*; Lgi, *Lottia gigantea*; Cte, *Capitella teleta*; Dme, *Drosophila melagonaster*; Cel, *Caenorhabditis elegans*; Nve, *Nematostella vectensis*; Tad, *Trichoplax adhaerens*.

Figure S5 – Conserved pairs using different cutoffs for the maximum number of intervening genes.

Relative number of total conserved pairs (including related and unrelated pairs/groups) in three representative species (human, black; Drosophila, red; and Nematostella, blue) using different cutoffs for the number of maximum intervening genes allowed (from 0 to 10; 4 was used in this study). For each species, all numbers are in relation to the number of pairs obtained for 10 intervening genes.

Table S1 – Conservation of microsynteny in the different species

This table shows the total number of conserved pairs for each species (pre-filtering), the number of conserved non-related gene pairs (post-filtering) and the expected number of pairs conserved by chance to ≥ 2 and ≥ 3 other lineages, based on gene order randomizations, and the corresponding false discovery rate (FDR) for the cutoffs in this study. The last column shows the total number of chromosomes/scaffolds in each genome assembly and the number of them that consists of a single gene (singletons).

Table S2 – Full dataset of microsyntenic pairs identified in this study.

This table contains different features for the 795 (column "i") types of pairs in all species where a type is present. These include: gene model/ids, number of intervening genes between the pair, total number of species in which it is conserved, orientation, fraction of orientation conservation ("Same-Ori"), most common orientation (per group "i"), whether the orientation is conserved, intergenic distance between the gene models, best D. melanogaster and human blast hits, type of pair ("OK", a simple pair of two genes whose orthology could be confidently assigned; "DEV CL", group of related *trans-dev* genes; "SUPER", groups of related genes whose orthology could not be confidently assigned), whether the pair of genes are related by sequences ("Paral/No"), the

sequence for both gene models in each species (separated by "XXX"), and data for tblastN searches in *A. queenslandica*, *C. owczarzaki*, *O. dioica* and *S. kowalevskii* (if positive hit: scaffold, orientation of the pair, minimum intergenic distance).

Abbreviations: Hsa, *Homo sapiens*; Gga, *Gallus gallus*; Xtr, *Xenopus tropicalis*; Cin, *Ciona intestinalis*; Odi, *O. dioica*, Bfl, *Branchiostoma floridae*; Spu, *Strongylocentrotus purpuratus*; Lgi, *Lottia gigantea*; Cte, *Capitella teleta*; Dme, *Drosophila melagonaster*; Cel, *Caenorhabditis elegans*; Nve, *Nematostella vectensis*; Tad, *Trichoplax adhaerens*; Aqu, *A. queenslandica*, Cow, *C. owczarzaki*.

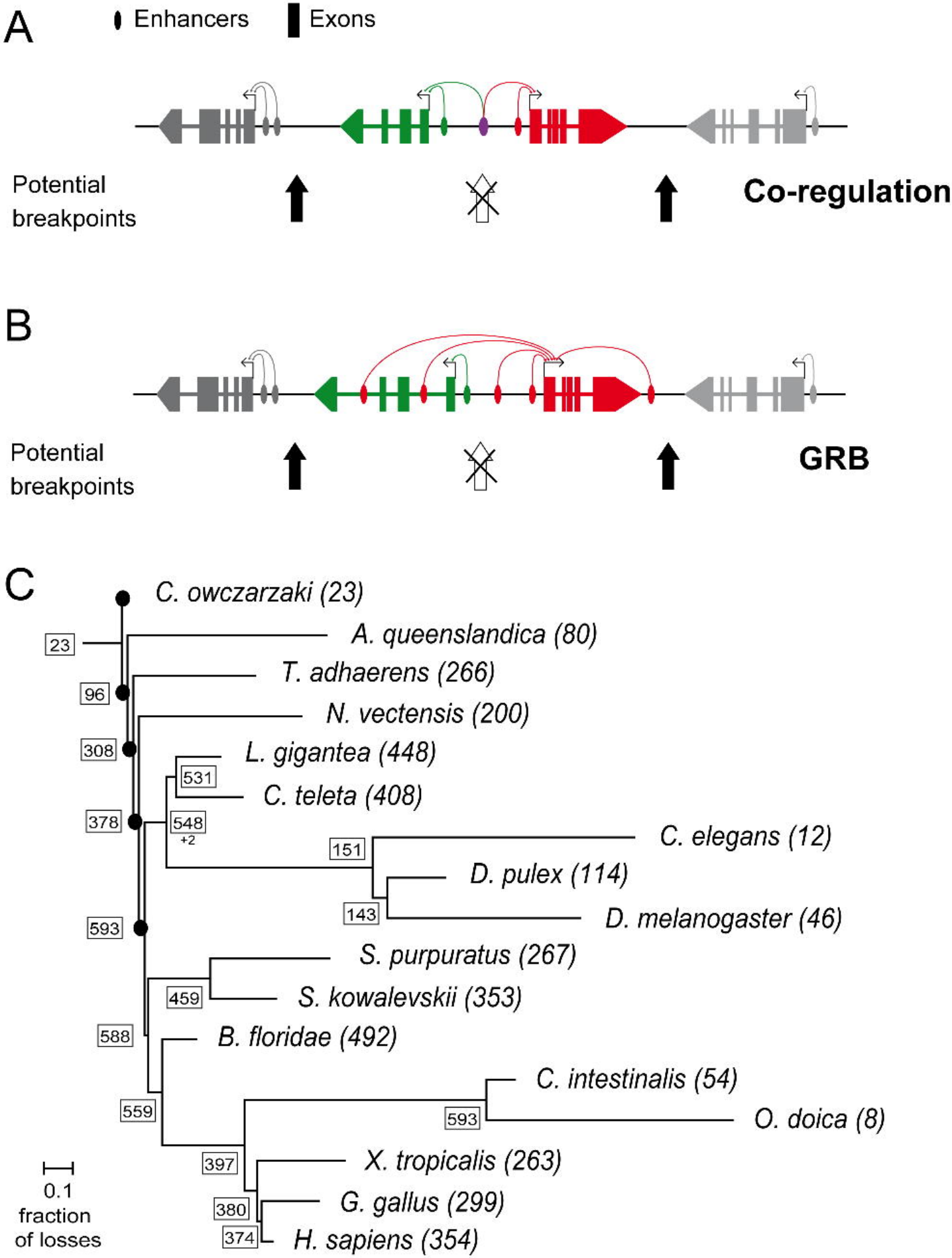
Table S3 – Analyses of pairs of paralogous genes

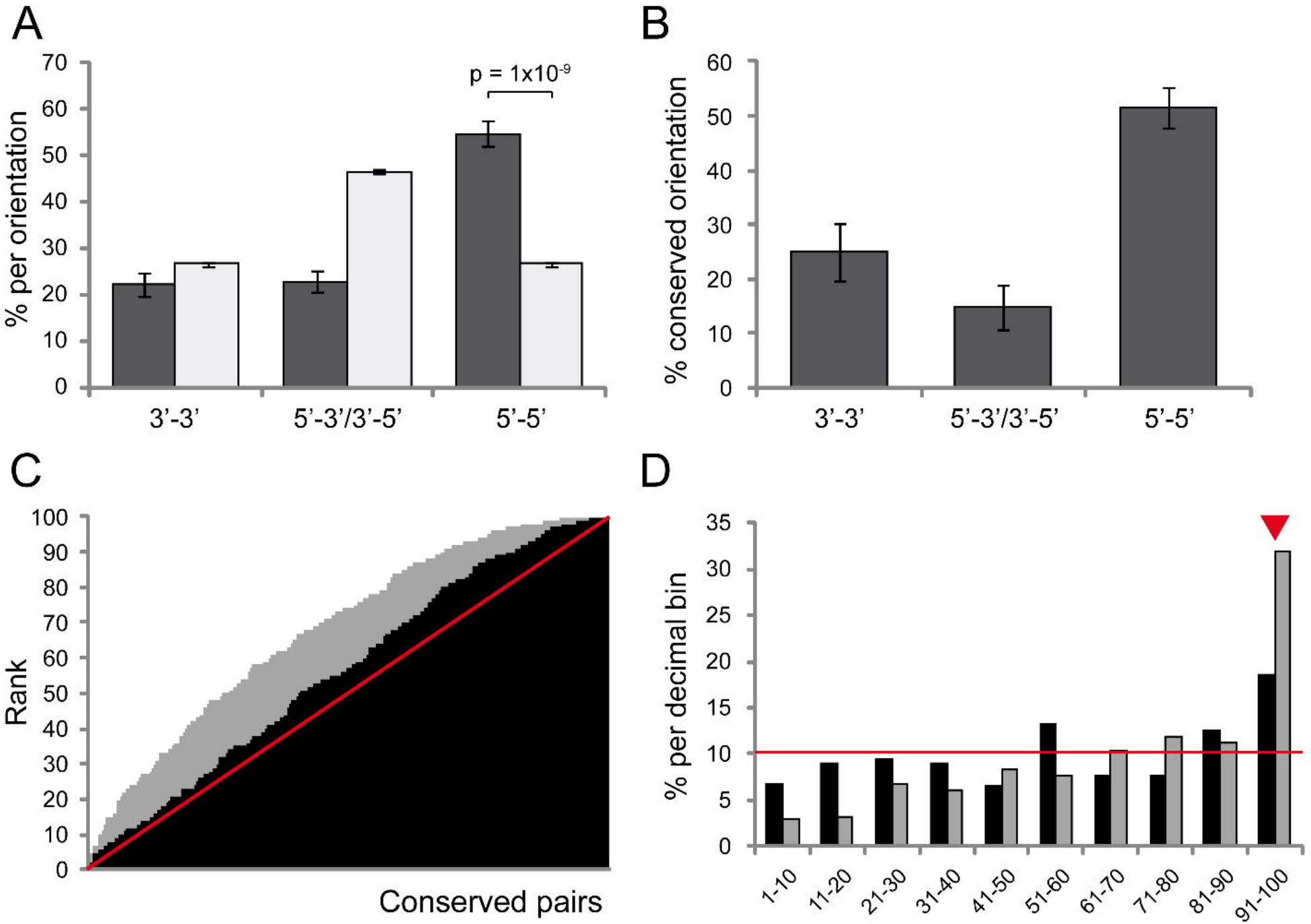
Each group of pairs of paralogous genes was analysed as described in Methods and the two defined scores (Apical Similarity (AS) and Group Score (GS)), and the corresponding p-values based on randomizations are given for each case. If both scores where significantly different than the random controls, the pair was annotated as likely to have been originated independently in each species by tandem duplications ("INDEPENDENT DUP").

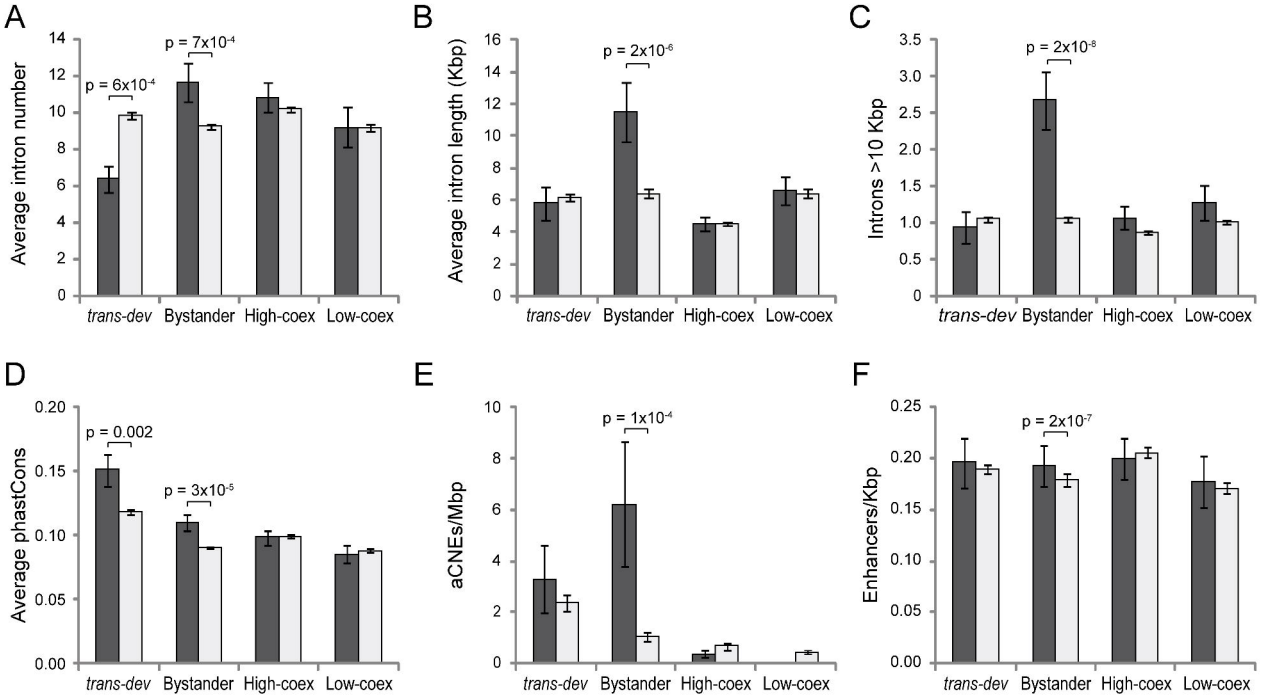
Table S4 – Putative human GRBs identified in this study

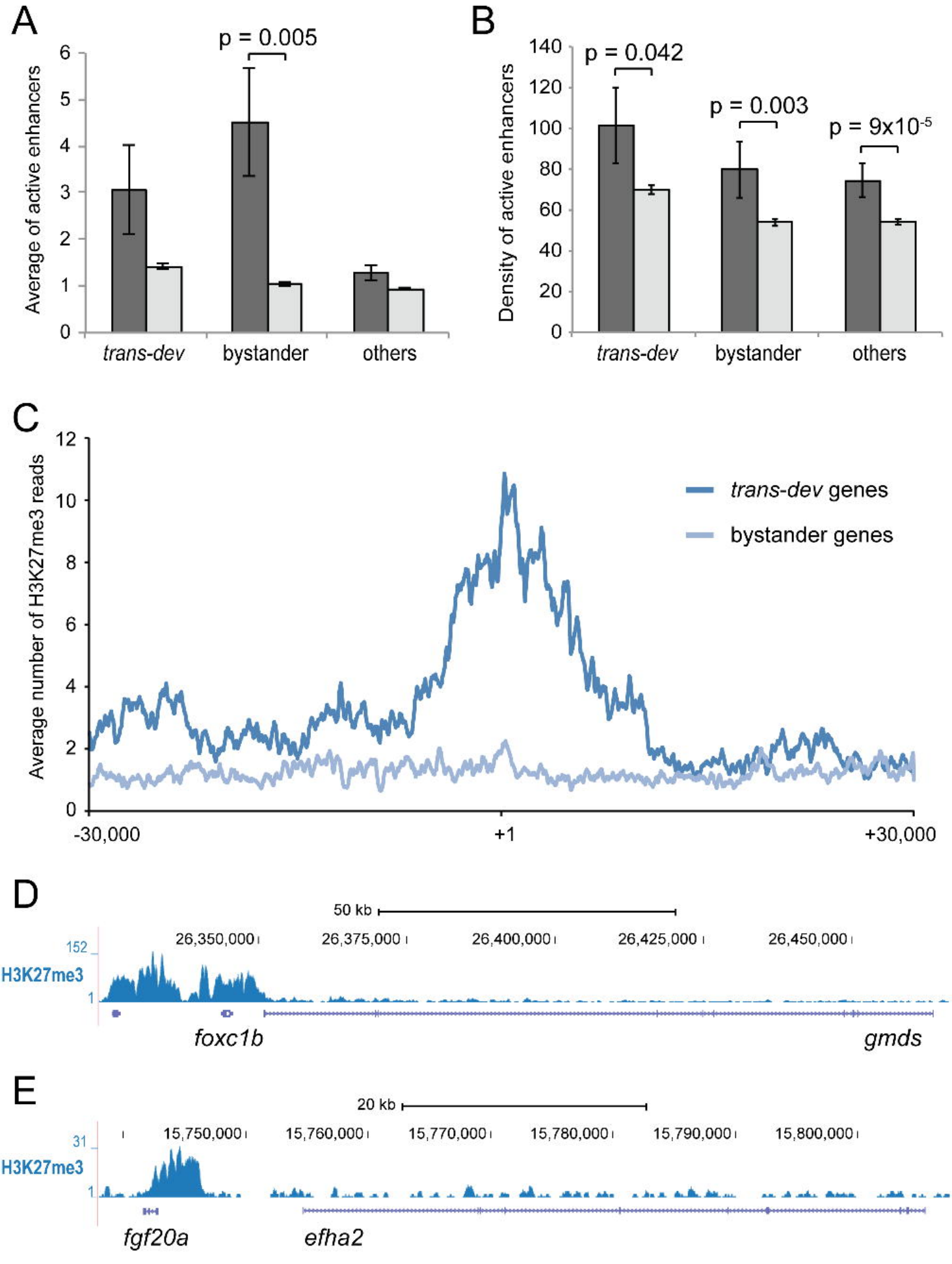
This table shows the pairs formed by a *trans-dev* and a non-developmental gene, as defined by GO categories. For each pair, the following features are given: number of intervening genes, total number of species in which the pair is present ("cons #sp"), orientation and whether it is conserved ("OriCONS"), intergenic distance, correlation of expression between the pairs ("CorrMAX"), whether the first or the second gene is the *trans-dev* (DEV) and different intron features for each of the genes (intron number, intron average, and number of introns longer than 10, 50 or 100Kbp).

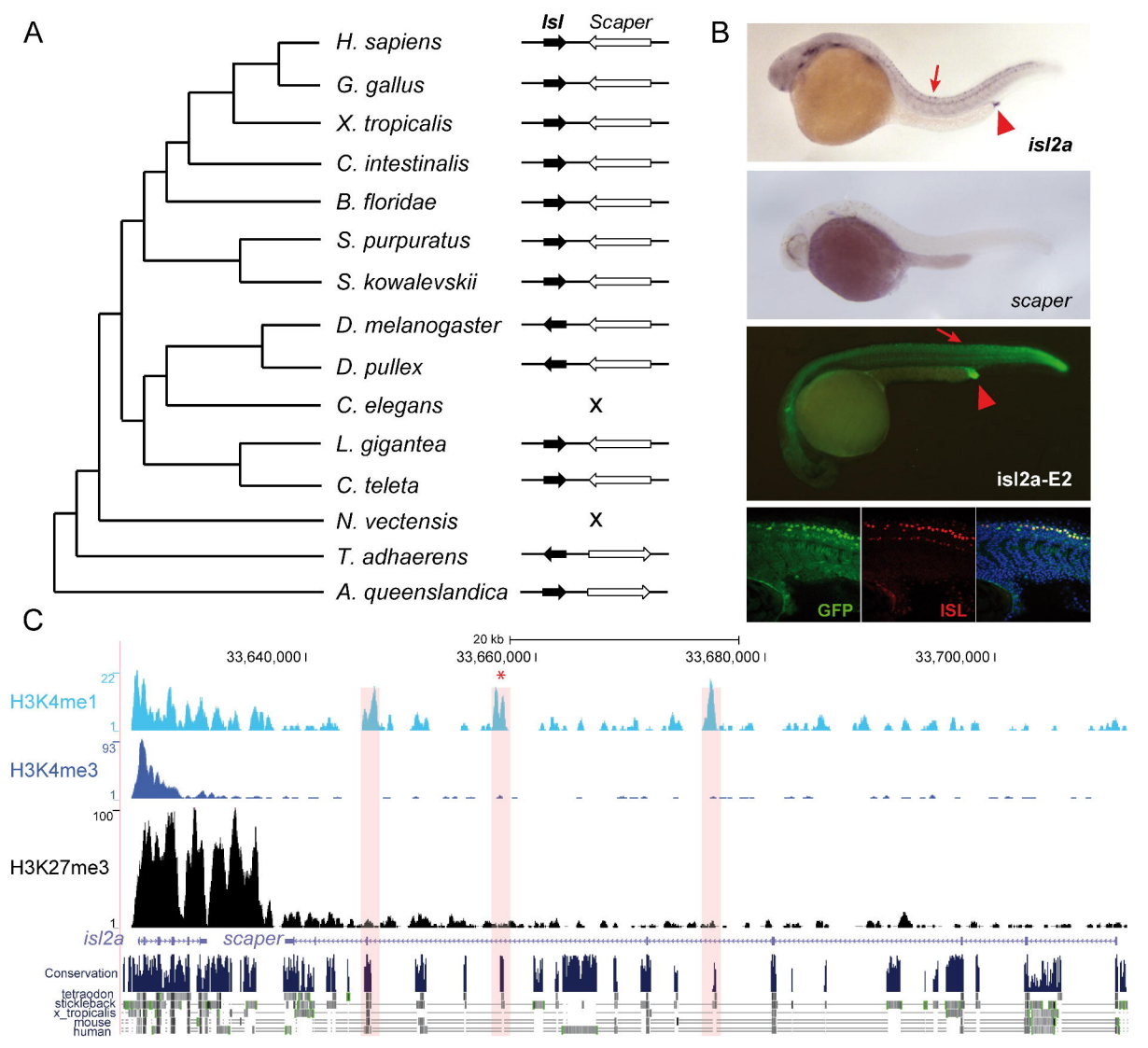
Table S5 – Primer sequences in this study

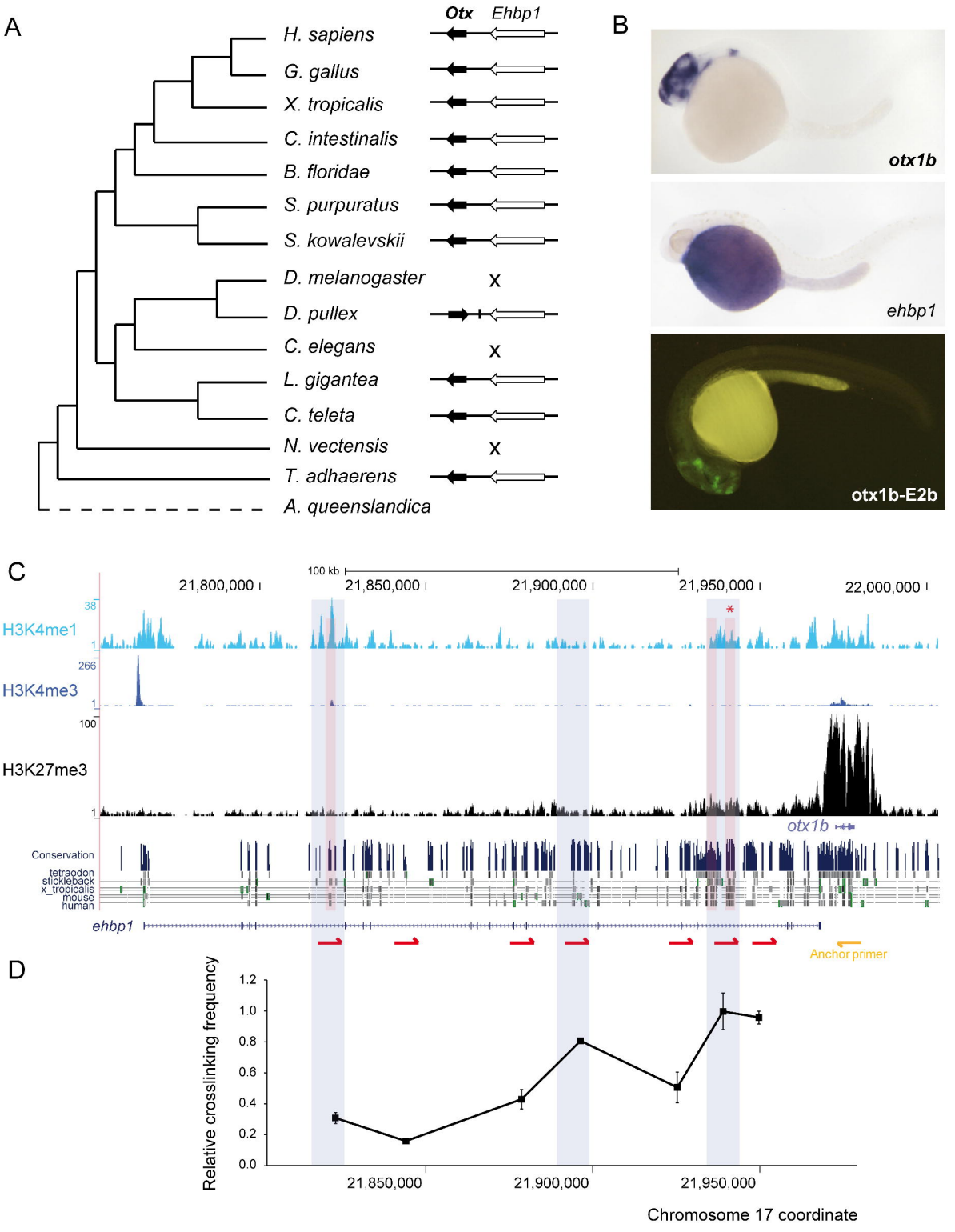

A
Enhancers
Exons
Potential breakpoints
Co-regulation
B
Potential breakpoints
GRB
C
C. owczarzaki (23)
A. queenslandica (80)
T. adhaerens (266)
N. vectensis (200)
L. gigantea (448)
C. teleta (408)
C. elegans (12)
D. pulex (114)
D. melanogaster (46)
S. purpuratus (267)
S. kowalevskii (353)
B. floridae (492)
C. intestinalis (54)
O. doica (8)
X. tropicalis (263)
G. gallus (299)
H. sapiens (354)
23
96
308
531
378
548
+2
151
143
593
459
588
559
593
397
380
374
0.1
fraction
of losses


A
% per orientation
70
60
50
40
30
20
10
0
p = 1x10-9
3'-3'
5'-3'/3'-5'
5'-5'
B
% conserved orientation
60
50
40
30
20
10
0
3'-3'
5'-3'/3'-5'
5'-5'
C
Rank
100
90
80
70
60
50
40
30
20
10
0
Conserved pairs
D
% per decimal bin
35
30
25
20
15
10
5
0
1-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100


A
Average intron number
p = 6x10-4
p = 7x10-4
B
Average intron length (Kbp)
p = 2x10-6
C
Introns >10 Kbp
p = 2x10-8
D
Average phastCons
p = 0.002
p = 3x10-5
E
aCNEs/Mbp
p = 1x10-4
F
Enhancers/Kbp
p = 2x10-7
trans-dev
Bystander
High-coex
Low-coex


A
p = 0.005
Average of active enhancers
0
1
2
3
4
5
6
trans-dev
bystander
others
B
p = 0.042
p = 0.003
p = 9x10-5
Density of active enhancers
0
20
40
60
80
100
120
140
trans-dev
bystander
others
C
Average number of H3K27me3 reads
0
2
4
6
8
10
12
trans-dev genes
bystander genes
-30,000
+1
+30,000
D
50 kb
26,350,000
26,375,000
26,400,000
26,425,000
26,450,000
152
H3K27me3
1
foxc1b
gmds
E
20 kb
15,750,000
15,760,000
15,770,000
15,780,000
15,790,000
15,800,000
31
H3K27me3
1
fgf20a
efha2


A
Isl
Scaper
H. sapiens
G. gallus
X. tropicalis
C. intestinalis
B. floridae
S. purpuratus
S. kowalevskii
D. melanogaster
D. pullex
C. elegans
L. gigantea
C. teleta
N. vectensis
T. adhaerens
A. queenslandica
X
X
B
isl2a
scaper
isl2a-E2
GFP
ISL
C
20 kb
33,640,000
33,660,000
33,680,000
33,700,000
H3K4me1
H3K4me3
H3K27me3
isl2a
scaper
Conservation
tetraodon
stickleback
x_tropicalis
mouse
human


A
Otx
Ehbp1
H. sapiens
G. gallus
X. tropicalis
C. intestinalis
B. floridae
S. purpuratus
S. kowalevskii
D. melanogaster
D. pullex
C. elegans
L. gigantea
C. teleta
N. vectensis
T. adhaerens
A. queenslandica
B
otx1b
ehbp1
otx1b-E2b
C
100 kb
21,800,000
21,850,000
21,900,000
21,950,000
22,000,000
H3K4me1
H3K4me3
H3K27me3
otx1b
Conservation
tetraodon
stickleback
x_tropicalis
mouse
human
ehbp1
Anchor primer
D
Relative crosslinking frequency
1.2
1.0
0.8
0.6
0.4
0.2
0.0
21,850,000
21,900,000
21,950,000
Chromosome 17 coordinate