PAPER I

Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates

Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates

Hiroshi Kikuta,[1] Mary Laplante,[1] Pavla Navratilova,[1] Anna Z. Komisarczuk,[1] Pär G. Engström,[2,3] David Fredman,[2] Altuna Akalin,[2] Mario Caccamo,[4] Ian Sealy,[4] Kerstin Howe,[4] Julien Ghislain,[5] Guillaume Pezeron,[5] Philippe Mourrain,[4] Staale Ellingsen,[1,10] Andrew C. Oates,[6] Christine Thisse,[7] Bernard Thisse,[7] Isabelle Foucher,[8] Birgit Adolf,[9] Andrea Geling,[9,11] Boris Lenhard,[1,2,12] and Thomas S. Becker[1,13]

[1]Sars Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway; [2]Computational Biology Unit, University of Bergen, 5008 Bergen, Norway; [3]Programme for Genomics and Bioinformatics, Department of Cell and Molecular Biology, Karolinska Institutet, 17177 Stockholm, Sweden; [4]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom; [5]Biologie Moléculaire du Développement, INSERM U368, Ecole Normale Supérieure, Paris, 75230 Paris, Cedex 05 France; [6]Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany; [7]IGBMC, CNRS/INSERM/ULP, BP10142, 67404 Illkirch, Cedex, France; [8]Unité de Génétique des Déficits Sensoriels, Institut Pasteur, F-75724 Paris Cedex 15, France; [9]Institute of Developmental Genetics, GSF Research Center, 85764 Neuherberg, Germany

We report evidence for a mechanism for the maintenance of long-range conserved synteny across vertebrate genomes. We found the largest mammal-teleost conserved chromosomal segments to be spanned by highly conserved noncoding elements (HCNEs), their developmental regulatory target genes, and phylogenetically and functionally unrelated "bystander" genes. Bystander genes are not specifically under the control of the regulatory elements that drive the target genes and are expressed in patterns that are different from those of the target genes. Reporter insertions distal to zebrafish developmental regulatory genes *pax6.1/2*, *rx3*, *id1*, and *fgf8* and miRNA genes *mirn9–1* and *mirn9–5* recapitulate the expression patterns of these genes even if located inside or beyond bystander genes, suggesting that the regulatory domain of a developmental regulatory gene can extend into and beyond adjacent transcriptional units. We termed these chromosomal segments genomic regulatory blocks (GRBs). After whole genome duplication in teleosts, GRBs, including HCNEs and target genes, were often maintained in both copies, while bystander genes were typically lost from one GRB, strongly suggesting that evolutionary pressure acts to keep the single-copy GRBs of higher vertebrates intact. We show that loss of bystander genes and other mutational events suffered by duplicated GRBs in teleost genomes permits target gene identification and HCNE/target gene assignment. These findings explain the absence of evolutionary breakpoints from large vertebrate chromosomal segments and will aid in the recognition of position effect mutations within human GRBs.

[Supplemental material is available online at www.genome.org.]

Conserved synteny, the maintenance of gene linkage on chromosomes of different species, is a prominent feature of vertebrate genomes (Ohno 1973). It is generally thought that large blocks of conserved synteny are relics of not yet occurred evolutionary chromosomal rearrangements (Nadeau and Taylor 1984). Obvious exceptions to this view are very large genes (up to 2.5 Mb in human) or the coregulated *hox* clusters (~5 Mb in human), including adjacent unrelated genes (for example, see Spitz et al. 2003; Lee et al. 2006). With the availability of multiple sequenced vertebrate genomes, however, it becomes evident that there are many additional genomic regions that are conserved to the extent of exact gene order. While the degree of conserved synteny is high between closely related species such as human and chimp, certain syntenic relationships are conserved from human to fish genomes (Boffelli et al. 2004). For human and mouse, Pevzner and Tesler (2003) postulated that there are numerous short "solid" synteny blocks flanked by "fragile" regions containing evidence for multiple ancient evolutionary breaks. While this analysis contradicts the long accepted random breakage model of chromosome evolution (Nadeau and Taylor 1984; Sankoff and Trinh 2005), it received further support by a recent publication suggesting that only a model assuming conserved blocks of genes plus extensive regulatory regions could accommodate the observed frequency of evolutionary breakpoint reuse (Peng et al. 2006).

It was shown recently that certain genomic regions contain arrays of highly conserved noncoding elements (HCNEs) clustered around developmental regulatory genes (Sandelin et al. 2004; Woolfe et al. 2005). These sequences are presumed to have *cis*-regulatory function, and the majority of those already tested

Present addresses: [10]Molecular Biology Institute, University of Bergen, 5008 Bergen, Norway; [11]Mischterlich and Partners, 80066 Munich, Germany.
[12]For information on bioinformatic methods, e-mail boris.lenhard@bccs.uib.no.
[13]Corresponding author.
E-mail tom.becker@sars.no; fax 47-55584305.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6086307.

have been shown to act as enhancers in transgenic reporter assays (for example, see de la Calle-Mustienes et al. 2005; Loots et al. 2005; Woolfe et al. 2005; Jeong et al. 2006). While not all regulatory sequences are recognizably conserved between human and teleost genomes, function can nevertheless be retained (Fisher et al. 2006). The activity of *cis*-regulatory elements regardless of conservation can be demonstrated through enhancer detection, the insertion of reporter-bearing vectors into the genomes of plants or animals (Sundaresan et al. 1995; Bellen 1999; Ellingsen et al. 2005).

In *Drosophila*, the majority of enhancer detector insertions were found within 200 bp of the transcription start site of the gene whose pattern is detected (Bellen et al. 2004), while in zebrafish, at least 20% of the expressing reporter insertions were found more than 15 kb away from the next transcriptional unit (Ellingsen et al. 2005). We propose here that regions detected by this approach, and by extension those found through bioinformatics approaches (Sandelin et al. 2004; Ahituv et al. 2005; Woolfe et al. 2005), contain long-range *cis*-regulatory elements distributed over large areas in and around their target genes and surrounding phylogenetically and functionally unrelated "bystander" genes, forming regions of conserved synteny we termed genomic regulatory blocks (GRBs). A bystander gene in this context is a gene that is not specifically under the control of the enhancers that define the GRB in which the bystander gene is located. Single-copy GRBs are protected from chromosomal breakage, while in cases of teleost duplication of GRBs, bystander genes functionally unrelated to the regulatory gene dominating the GRB have often been lost by neutral evolution, a phenomenon predicted by the duplication degeneration complementation model (Force et al. 1999). As we show here, the combination of human/teleost synteny, enhancer detection, and GRB duplication analysis allows recognition of target versus bystander genes and permits annotation of HCNEs to target genes within a minimal conserved syntenic chromosomal segment. Finally, analysis of duplicated teleost GRBs can identify nonduplicated interlocked bystander genes as probable false candidates in the mapping of human disease mutations.

Results

Genome-wide properties of largest syntenic blocks

We devised a rule-based procedure to estimate minimal blocks of synteny between human and zebrafish (see Methods). The distribution of the genomic spans of the resulting synteny blocks is shown in Figure 1A, with the position of blocks harboring the gene loci analyzed in this paper labeled by their inferred target gene. It is obvious that the studied blocks are among the longest ones detectable. In addition, we compared the distributions of synteny block spans for different functional categories of genes. This comparison shows that genes encoding developmental transcriptional regulators tend to be surrounded by larger regions of synteny than other functional categories of genes ($P < 10^{-6}$) (Fig. 1B). In addition, we have examined the 100 largest synteny blocks in a zebrafish/human comparison and detect a developmental regulatory gene and associated HCNEs in almost every one of them (Supplemental Table S1). It has been suggested that regulatory elements residing in adjacent genes constitute a mechanism to conserve synteny (MacKenzie et al. 2004; Ahituv

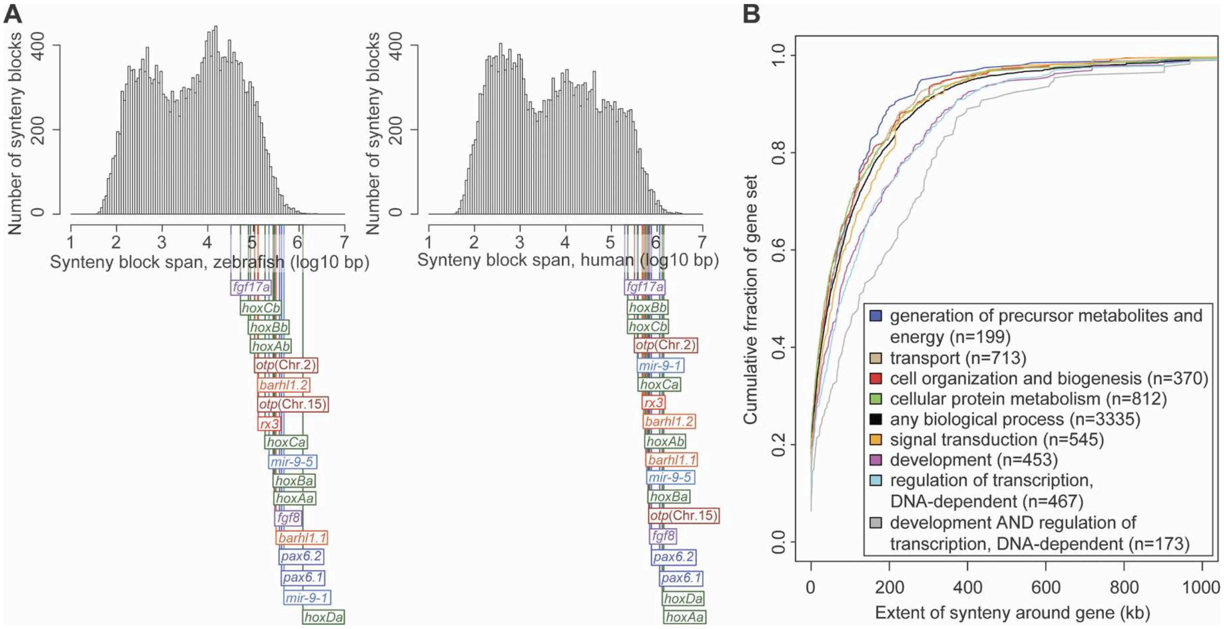

Figure 1. The studied loci are within large synteny blocks. (*A*) Histograms show the span of synteny blocks in the zebrafish (*left*) and human (*right*) genomes. Colored lines indicate the genomic spans of synteny blocks for the loci investigated in this study and, for comparison, the loci of the seven zebrafish *hox* clusters. Note that zebrafish gene symbols are used in both histograms in order to differentiate between synteny blocks that overlap on the human genome (e.g., the *pax6.1* and *pax6.2* synteny blocks, which partially overlap at the human *PAX6* locus). Synteny blocks were computed based on alignments between the two genomes as described in Methods. (*B*) Each curve shows the cumulative distribution of extent of synteny around genes participating in a particular biological process. The distributions for the processes transcriptional regulation and development grow significantly slower compared to any of the other processes investigated ($P < 0.004$; one-tailed Kolmogorov-Smirnov test). For genes involved in both of these processes, the difference is highly significant ($P < 1 \times 10^{-6}$). Numbers within parentheses indicate the number of genes annotated to a process and located within a synteny block.

et al. 2005; Goode et al. 2005; Gomez-Skarmeta et al. 2006; McEwen et al. 2006; Vavouri et al. 2006).

Blocks of synteny duplicated in teleost genomes define GRBs

On the assumption that synteny blocks containing developmental regulatory genes are kept together by essential regulatory elements, we mapped HCNEs and genes within areas in which both copies of the presumed GRB target gene were retained after whole-genome duplication in teleosts. Such duplicated loci are assumed to share the expression domains of the ancestral single locus (Force et al. 1999). If evolutionary constraint acted on an area through long-range regulatory elements, resulting in an extended conserved domain including neighboring genes, this constraint might be expected to be relaxed upon duplication, and changes might be detected in both gene retention and the preservation of HCNEs after GRB duplication.

Orthopedia (*otp*), a homeobox gene expressed in the mouse hypothalamus, is necessary for cell migration, proliferation, and differentiation (Acampora et al. 1999). The human *OTP* locus (Fig. 2A) is spanned by a large array of HCNEs. Many of the distal HCNEs upstream of the gene are located in introns of a neighboring gene, *AP3B1*. *OTP* has two zebrafish orthologs, *otp*, expressed in hypothalamus and hindbrain (Supplemental Fig. S1), and *XP_683186.1*. The latter has lost the entire distal upstream part of the HCNE array, while retaining many of the proximal and intronic ones. The other copy, *otp*, has retained the HCNEs that inhabit *AP3B1* introns in human, but the *AP3B1* exons have been lost. There is a single copy of *ap3b1* elsewhere in the zebrafish genome, which has almost no intron sequence conservation with the human gene. This suggests a mechanism whereby a duplicated GRB can selectively retain a subset of regulatory inputs and lose others, either by accumulation of mutations or by a chromosome break that removes a part of the HCNE array from one copy of the GRB together with any bystander genes. This is a plausible explanation for what happened to *ap3b1* after *otp* GRB duplication: The entire interval around this gene broke off from one copy of the GRB and landed elsewhere in the genome. The break was not selected against because the other copy of *otp* still had all the regulatory inputs in place. Once detached from their target gene, the HCNEs in the introns of *ap3b1* disappeared by neutral evolution. In the other copy (*otp*), the opposite happened: while the HCNE array was retained, the *ap3b1* bystander gene that originally harbored them was lost.

Human *BARHL1*, a homeobox transcription factor gene (Bulfone et al. 2000), has two zebrafish orthologs, *barhl1.1* and *barhl1.2* (Fig. 2B) (Colombo et al. 2006). *BARHL1* is spanned by an array of HCNEs; in zebrafish, some HCNEs are present in only one of the two duplicated loci. The mammalian syntenic block contains seven known genes, out of which two (*BARHL1* and *TSC1*) have been retained in both copies. Of the four genes between them, three were retained at the *barhl1.2* locus only. The fourth one (*GTF3C4*) was retained only at the *barhl1.1* locus, flanked with two HCNE-containing gene desert-like regions. At the human locus, these HCNEs are located within introns of *DDX31* and C9orf98, and some of them are also found within orthologous introns at the *barhl1.2* locus. The observed disentangling of HCNEs and genes in zebrafish suggests that the four human genes (and their zebrafish orthologs) are unrelated to the HCNEs with which they are nested. We can therefore label them bystander genes. In contrast, if both copies of a gene have been kept, as with the zebrafish orthologs of *BARHL1* and *TSC1*, then no prediction can be made through genome inspection alone on the specificity of the HCNEs with respect to either gene, and both might consequently be regulated by these elements or represent overlapping GRBs.

The human growth factor gene *FGF8* is in a synteny block with the downstream gene *FBXW4* throughout chordates, conserved even in *Ciona* genomes (data not shown). This syntenic block is inverted in all teleosts relative to mammalian genomes and has undergone duplication in teleost genomes (Fig. 3A). The zebrafish duplicate maps to chromosome 1 and is annotated as *fgf17a*. This block has retained *NP_056263.1* and *POLL*, two genes that in the human genome are downstream from *FGF8*, but has undergone deletion of *fbxw4*. Even though this gene was originally annotated as *fgf17* (Reifers et al. 2000), it is more similar to *fgf8* in sequence and expression pattern (for expression patterns of *fgf8*, *fbxw4*, *fgf17a*, and *poll*, see Supplemental Fig. S1). We

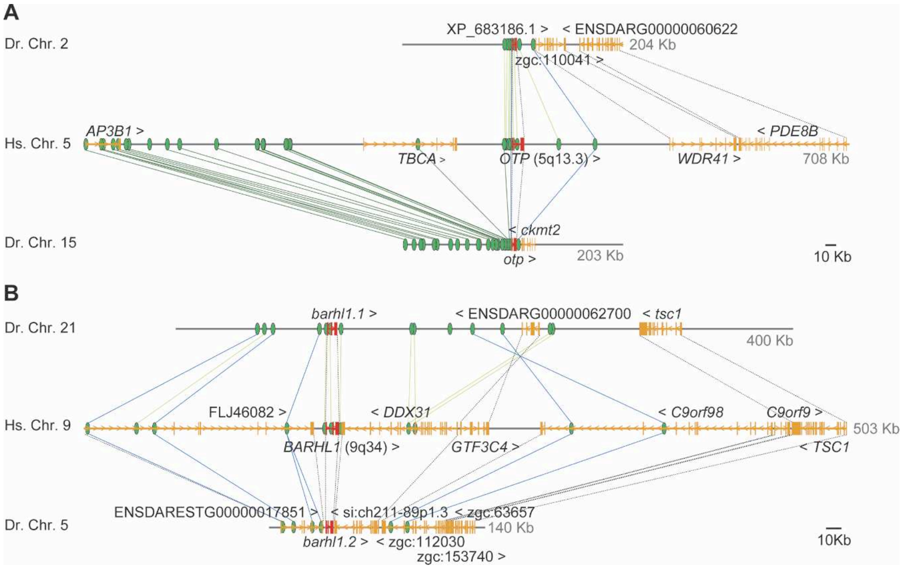

Figure 2. Duplicated zebrafish GRBs. (*A*) Orthopedia (*otp*). (Green ovals) HCNEs; (orange and red gene models) bystander and target genes, respectively. After duplication, one of the zebrafish GRBs (*upper* track) lost most of its upstream parts, including a large HCNE array and *AP3B1* and *TBCA* genes, and *AP3B1* and *TBCA* landed elsewhere in the genome (data not shown), while the other (*lower* track) kept most of the HCNE array while *AP3B1* and *TBCA* were lost by neutral evolution. Cross-species sequence comparisons indicate that the region upstream of *ckmt2* harbors an *RASGRF2* ortholog (data not shown) for which no full-length cDNA is available. (*B*) The duplicated *barhl1* loci. *Barhl1.1* (*upper* track) lost the bystander genes *DDX31*, C9orf98, and C9orf9, but retained all of the HCNEs found in the human locus, and there is evidence of an inversion that occurred between *barhl1.1* and *tsc1* (crossed lines interconnecting HCNEs), which includes the only retained copy of the *GTF3C4* gene. *Barhl1.2* (*lower* track) lost the downstream *GTF3C4*, but retained all other annotated genes and some HCNEs.

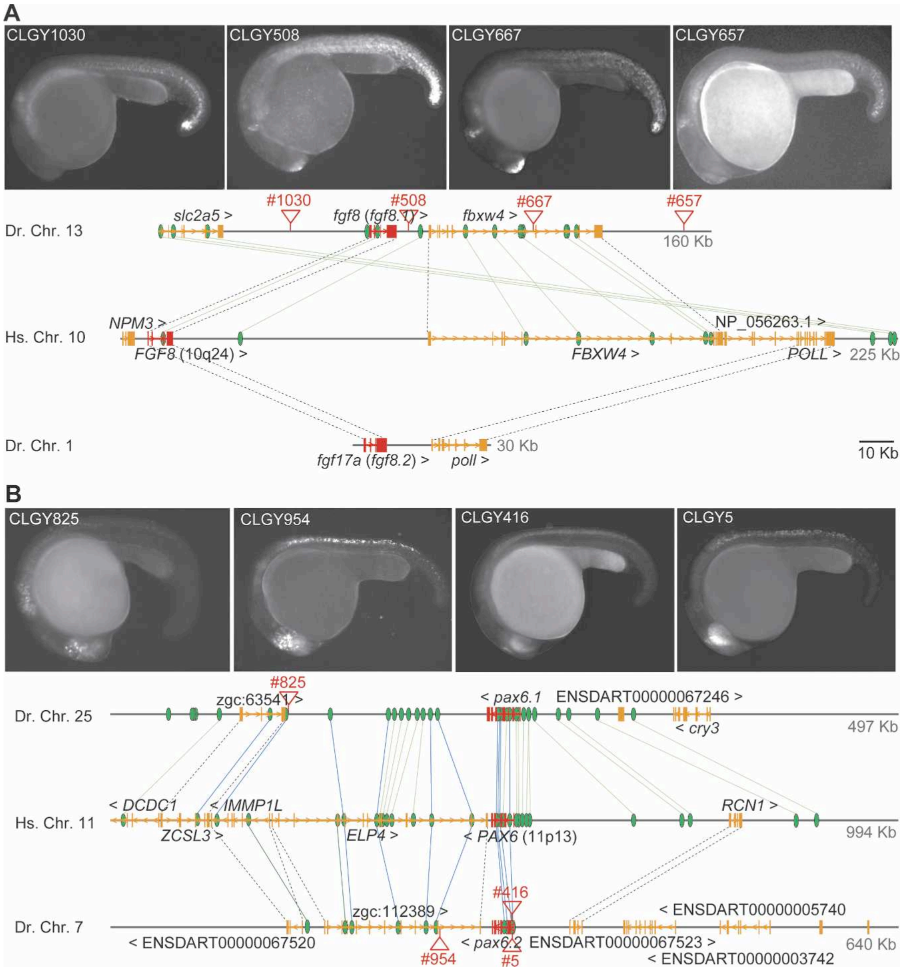

Figure 3. (*A*) The duplicated *fgf8* loci of zebrafish. (*Upper* track) The current *fgf8* locus has kept its downstream neighbor *fbxw4* but has lost the farther downstream NP_056263.1 and *POLL* genes, which are, however, retained downstream from *fgf17a*. NP_056263 is conserved in zebrafish at this position, but currently not annotated (data not shown). Both duplicated loci have lost neighboring genes, even though the syntenic relationship of human *NPM3*, *FGF8*, and *FBXW4* is conserved across chordate genomes. HCNEs located in the upstream gene *slc2a5* are found far downstream from *FGF8* in human, with no evidence of a human counterpart of *slc2a5*. Numerous HCNEs were also found within *FGF8*, downstream from the gene, and inside the bystander *FBXW4*, and most of them are conserved only in zebrafish *fgf8*. Enhancer detection insertions (red triangles) cover the entire GRB and show partial (CLGY1030) or complete *fgf8* expression patterns at 24 h. In these and all other images, anterior is to the *left* and dorsal to the *top*. (*B*) Human *PAX6* GRB covers 1 Mb, containing five bystander genes. One of these, the far downstream *DCDC1* is not found in teleost genomes; of the others, only one copy has been retained either in *pax6.1* or *pax6.2* GRBs. Insertion CLGY825 in the *pax6.1* GRB shows the correct expression pattern (cf. Supplemental Fig. S1) despite being adjacent to the *ZCSL3* ortholog, while CLGY954 in the *pax6.2* GRB is inside *elp4* yet has the expression pattern of *pax6.2* (cf. CLGY5 and −416 inside *pax6.2* and *pax6.2* expression pattern in Supplemental Fig. S1).

propose that this gene should be annotated as *fgf8.2*, and the current *fgf8* as *fgf8.1*.

Enhancer detection allows visualization of GRB regulatory content

Using enhancer detection in zebrafish (Ellingsen et al. 2005), we isolated insertions inside GRBs, which independently verify these regions as having unique *cis*-regulatory content. We recovered four insertions in the *fgf8* GRB on chromosome 13, and all of them display an *fgf8*-like pattern (Fig. 3A; Supplemental Fig. S1).

One insertion (CLGY1030; position 33,797,961) was located ~29 kb upstream of the *fgf8* start codon, and one (CLGY508; position 33,831,088) 4684 bp downstream from the last exon of *fgf8* (Fig. 3A). One further insertion was mapped into intron 5 of *fbxw4* (CLGY667; position 33,869,215), and one (CLGY657; position 33,898,475) downstream from the last exon of *fbxw4* (Fig. 3A). The insertion 29 kb upstream of *fgf8* (CLGY1030) mimics the *fgf8* expression pattern only in the tail bud, while the three insertions downstream from *fgf8*, inside *fbxw4*, and downstream from *fbxw4* mimic a more complete *fgf8* expression pattern (telencephalon, optic stalk, mid-hindbrain boundary, somites, heart, olfactory pits, and tail bud; Fig. 3A). CLGY508, the insertion closest to *fgf8*, also shows expression in the apical ectodermal ridge (AER) in the pectoral fin bud domain. The organization of regulatory elements around *fgf8* has recently been assayed (Inoue et al. 2006), but our results suggest that there must be additional elements inside *fbxw4*. Hence, *fgf8* and *fbxw4* are part of the same GRB, and insertions over a 100-kb section of chromosome assume a partial or near complete *fgf8* expression pattern, while the expression of *fbxw4* is ubiquitous but weak (Supplemental Fig. S1).

Pax6 is a gene involved in vertebrate retinal and CNS development, and human *PAX6* is mutated in aniridia (Glaser et al. 1992; Jordan et al. 1992). The gene is duplicated in teleosts. We recovered two insertions on chromosome 7 in intron 3 of *pax6.2*, and one insertion ~68 kb downstream from the transcriptional unit, in an intron of the downstream gene *elp4* (Fig. 3B). These insertions show the expression pattern of *pax6.2*, suggesting that the *cis*-regulatory information driving *pax6.2* is available inside *elp4*, while *elp4* expression is much more widespread (Supplemental Fig. S1) and thus does not appear to be regulated specifically by the elements within its introns. We also mapped an insertion 116 kb downstream from *pax6.1* with the corresponding expression pattern (Fig. 3B; Supplemental Fig. S1). However, although the conservation of synteny with mammalian and avian genomes suggests that this genomic area was duplicated in its entirety, neither of the downstream genes *elp4* and *immp1l* was retained downstream from *pax6.1*. Thus, while the entire region of 400 kb has been conserved in both duplicates, the bystander genes *elp4* and *immp1l* were retained only in the *pax6.2* locus (Fig. 3B) and disappeared downstream from *pax6.1*, leaving behind a 120-kb gene desert spanned by multiple HCNEs. Gene deserts have been recognized as extended regions of regulatory activity that resist evolutionary chromosomal rearrangements (Ovcharenko et al. 2005); we propose that GRBs are functionally equivalent to gene deserts, the only difference being the absence of bystander genes in gene deserts, which in GRBs do not seem to affect, or be affected by, the long-range regulatory activity in the region.

Regulatory information is available in large areas around developmental regulatory genes

The zebrafish *id1* transcription factor gene (formerly *id6*) on chromosome 11 (chr11) bears 52% similarity with human *ID1*. Despite the relatively low conservation at the protein level, there are two HCNEs conserved between human and zebrafish (Fig. 4A), suggesting that the human gene and the zebrafish gene along with surrounding sequence share a common ancestor. When compared to the *Tetraodon* genome, >100 kb of the zebrafish *id1* locus aligns with ~30 kb of the *Tetraodon id1* locus, including multiple HCNEs (Fig. 4A). We mapped nine insertions in an area of ~50 kb, eight upstream and one downstream of *id1*, and all show virtually identical global expression patterns, highly similar to that of *id1* (Supplemental Fig. S1), although there may be small-scale differences (Fig. 4A). Thus, within a large area, *cis*-regulatory information is driving inserted enhancer detection vectors in highly similar expression patterns, largely independent of insertion location.

Syntenic blocks of multiple genes may contain regulatory information for a single developmental regulatory gene

Rax is a vertebrate homeobox gene essential for retinal development (Mathers et al. 1997; Voronina et al. 2004; Stigloher et al. 2006). The zebrafish ortholog *rx3* is located on Chromosome 21 within an extended region of conserved synteny compared with the human *RAX* locus. There are two unrelated genes located upstream of *RAX* in this synteny block, *CPLX4* and *LMAN1*. We mapped an insertion within intron 7 of zebrafish *lman1*, ~38 kb upstream of *rx3*, and the insertion mimics the expression pattern of *rx3* (Fig. 4B). Even though in zebrafish/human alignments of the *rx3/RAX* genomic neighborhood only a single HCNE exceeds the threshold we applied for genome-wide detection of HCNEs in this work (Fig. 4B), multiple elements are found in zebrafish/*Tetraodon*/fugu *rx3* alignments, some of which are within the introns of the neighboring genes (data not shown). These findings suggest that, although several genes are found within this block of conserved synteny, the HCNEs in the region are functional regulatory elements acting on *rx3* and that the spatial relationship of genes must be conserved, even though many of the regulatory elements are not recognizably conserved between human and teleost genomes. Long-range enhancer detection therefore provides an experimental means of identifying target and bystander genes. In this case, the bystander genes *cplx4* and *lman1* have much broader expression patterns than *rx3* and are not under specific regulation of *rx3* regulatory elements.

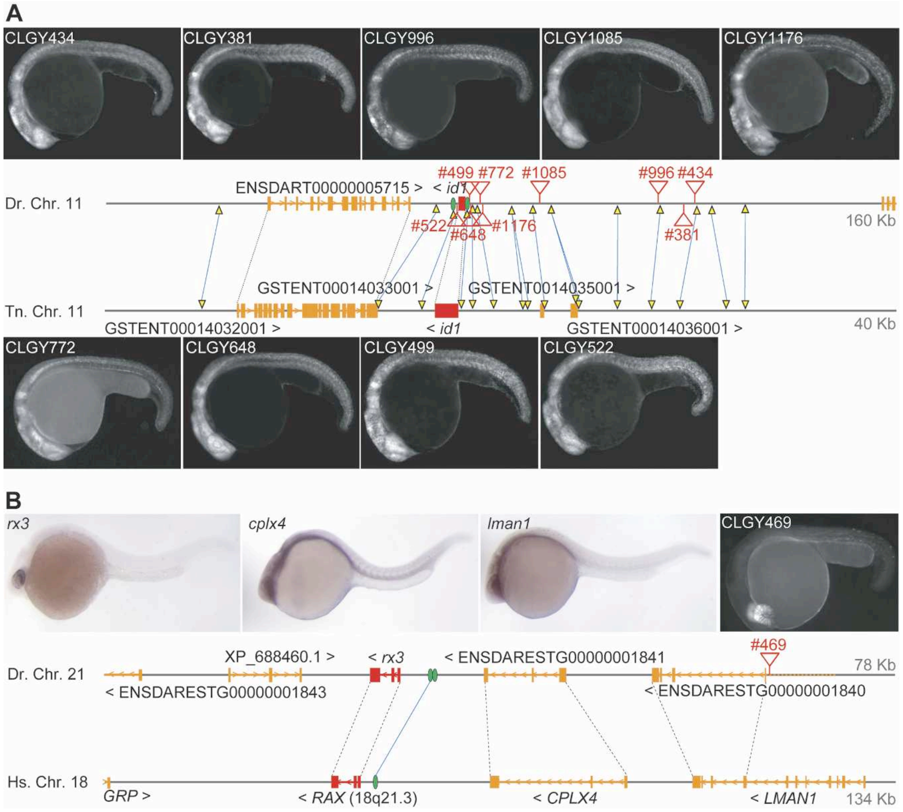

Figure 4. Genomic regulatory blocks are extended regions of *cis*-regulatory content. (*A*) Zebrafish *id6* locus is orthologous to tetraodon *id1*, located in a gene desert spanned by conserved noncoding elements (yellow triangles connected with blue lines, corresponding to Ensembl translated BLAT alignments). Two HCNEs are conserved from fish to human *id1* loci (green ovals in zebrafish locus). The insertions recovered are labeled as red triangles with transgenic line numbers. Note the same global expression pattern regardless of insertion position (id6 expression pattern in Supplemental Fig. S1). (*B*) GRB of human *RAX* and zebrafish *rx3* loci, consisting of *RAX*, *CPLX4*, and *LMAN1* genes. An insertion within *lman1*, CLGY469, assumes the expression pattern of *rx3*, not of the gene it is inserted into. While only one HCNE can be discerned in this alignment, multiple elements are found in fish/fish alignments (data not shown).

MicroRNAs can be target genes in GRBs

In *Drosophila*, transcriptional control of some miRNAs is comparable to that of protein-coding genes (Biemar et al. 2005); additionally, REST-binding sites were shown to be involved in miRNA regulation in mammalian genomes (Conaco et al. 2006). Several miRNAs are hosted within other genes, but the majority appear to be transcribed from their own promoters. We recovered an insertion, on Chromosome 16, in a zebrafish homolog of transcriptional activator of the *c-fos* promoter (*C1orf61*), which also hosts *miRNA 9-1* (Fig. 5A). The expression, in the dorsal telencephalon, of the inserted vector is identical to that of *mirn9* at 24 h postfertilization (Wienholds et al. 2005) and is also identical to the expression pattern of zebrafish *c1orf61*, while the nearest downstream gene, *rhbg*, has a different expression pattern (Fig. 5A; Supplemental Fig. S1). Both genes are embedded in an area with conserved

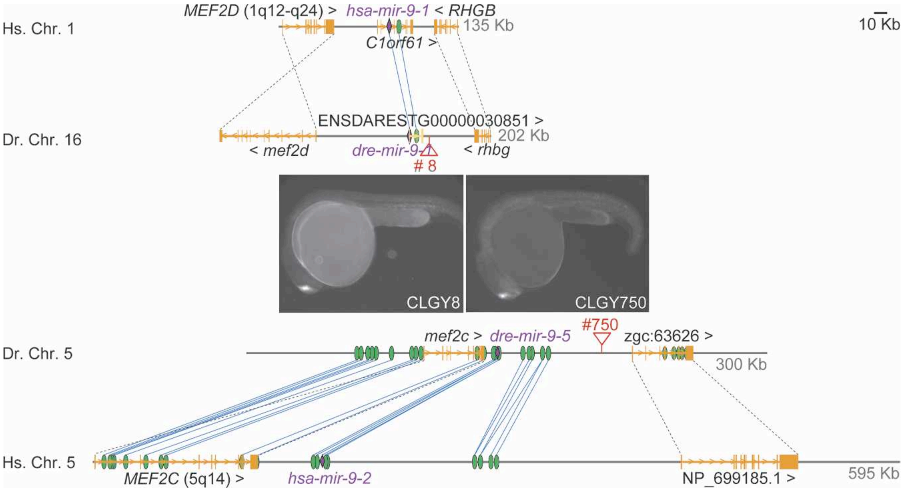

Figure 5. Two GRBs associated with genes encoding microRNAs; both also containing a myocyte enhancer factor gene. (*Top*) Human/zebrafish *mirn9-1* occupies the same relative position in both genomes, flanked by *rhbg* and *mef2d* genes, and hosted within *C1orf61* in human, also represented by an EST sequence in zebrafish (yellow). Insertion CLGY8 is expressed, at 24 hpf, in the dorsal telencephalon. (*Lower* track) Zebrafish *mirn9-5* and human *mirn9-2* are located near *mef2c* in a gene desert spanned by multiple HCNEs (green ovals). Insertion CLGY750 is ~100 kb distal to *mirn9-5* and, at this stage, has the same expression pattern as *mirn9-1*, in the dorsal telencephalon. For expression patterns of *zgc:63626*, *mef2c*, and *rhbg*, see Supplemental Figure S1.

synteny throughout vertebrates, and this interval also includes a gene encoding a myocyte enhancer factor 2d (*mef2d*).

A paralog of *mef2d*, *mef2c*, is found on Chromosome 5 upstream of a gene desert that also harbors an miRNA gene, encoding human *MIRN9-2*/zebrafish *mirn9-5*. An insertion within this gene desert, ~100 kb downstream from the miRNA, has an expression pattern also resembling *mirn9* (Fig. 5; Wienholds et al. 2005). In contrast, *mef2c* is expressed in somites and myotomes, and *zgc:63626* in a widespread pattern (Supplemental Fig. S1).

These results also show that miRNAs can be regulated by the same type of enhancers as are developmental regulatory genes, regardless of whether they are hosted in protein-coding genes or transcribed from their own promoters. Uncharacterized miRNAs that appear to be target genes in GRBs are prime candidates for the investigation of their role in development. The conservation of synteny in these blocks around miRNA target genes indicates that these extended regulatory domains may be sensitive to chromosomal rearrangements resulting in position effect mutations and possibly harbor human disease breakpoints.

Determination of HCNE density allows annotation of target genes within GRBs

Inspection of other clusters of HCNEs detected several of them spanning and most likely targeting miRNA gene loci in all vertebrates. This was enabled by the fact that the density of HCNEs within a GRB is not uniform (Fig. 6) and often peaks close to or within introns of the most likely target gene. This property is immediately applicable to the determination and annotation of targets in as-of-yet uncharacterized GRBs. It correctly points to miRNA genes as targets of HCNEs in the vicinity of *mef2c* genes (Fig. 5) as well as other miRNAs: We found one example of a cluster of miRNAs on human Chromosome 2, next to *EFEMP1*, an extracellular matrix protein implicated in retinal dystrophy (OMIM*601548; Stone et al. 1999). The zebrafish orthologs of these miRNAs, *mirn216* and *mirn217*, are both expressed in the retina (Wienholds et al. 2005). Another example is an miRNA cluster on human Chromosome 7, in between the protein-coding genes *NRF1* and *UBE2H*. This GRB contains at the center miRNAs *182*, *96*, and *183*. Most likely, the HCNEs found in this cluster regulate these miRNAs, not the adjacent coding genes, none of which falls into the functional categories of protein-coding genes typically targeted by GRBs (Sandelin et al. 2004).

The use of HCNE density to estimate the GRB target gene can be ambiguous in several cases in which two or more GRBs are

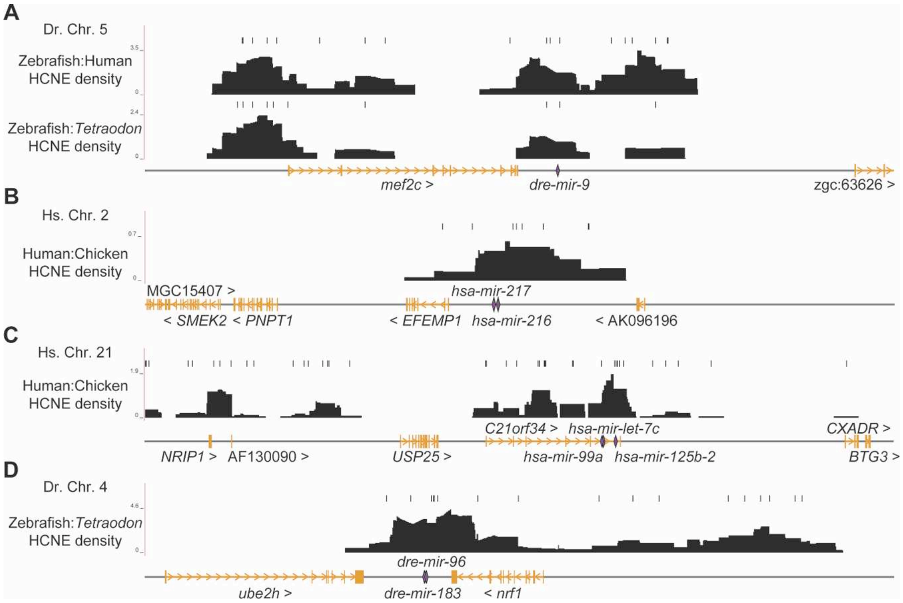

Figure 6. Density plots of HCNEs across human and zebrafish *miRNA* GRBs. HCNE density (black) often peaks near miRNAs (purple) in the vicinity of other genes (orange), suggesting miRNAs as HCNE target elements within GRBs. Density profiles calculated against species of different evolutionary distances separately identify the same hotspot for each region, but at different resolutions. The plots shown here are the most informative for each region. Our HCNE density score represents the number of bases within HCNEs determined by computational analysis (see Methods) divided by the number of non-exonic bases in sliding windows across zebrafish and human chromosomes (20-kb window, 100-bp step size and 100-kb window, 1-kb step size, respectively). The bars *above* density profiles represent HCNEs.

apparently adjacent. Of the cases described so far, the *fgf8* GRB in vertebrates might be fused with an adjacent GRB targeting the *LBX1* and/or *TLX1* genes. In this case, the density appears as if *fgf8* is at the tail of a large GRB (data not shown), and only the insertions described above and the fact that the ancestral *fgf8/17/18* and *fbxw4* orthologs colocalize in *Ciona* point to the fact that this is an evolutionary separate GRB.

Discussion

Conservation of human/teleost synteny is under evolutionary pressure

In this study, we show that long-range enhancers and their regulatory target genes inhabit chromosomal segments that often include bystander genes that are phylogenetically and functionally unrelated to the target gene. Since the cases we have shown here represent loci that conserve syntenic relationships through all vertebrate genomes, the target genes within these GRBs as well as their inferred *cis*-regulatory sequences are likely fundamental to general vertebrate development and ontogeny. The loci in this paper are not the first to be shown to be kept together by regulatory sequences: *hox* clusters are conserved throughout most metazoan genomes, as are other gene clusters, such as *irx*, as well as certain loci that consist of tandem duplications of regulatory genes, for example, *myf5/mrf4* and *dlx* genes (Zerucha and Ekker 2000; Carvajal et al. 2001; Spitz et al. 2003; de la Calle-Mustienes et al. 2005; Lee et al. 2006). In these cases, however, the proposed mechanism underlying conserved synteny is the coregulation of several genes by the same regulatory sequences. In contrast, the evolutionary mechanism we propose here is the interdigitation of regulatory sequences and their target gene with functionally and regulationally unrelated bystander genes. This constraint is temporarily relaxed through duplication of these blocks (Fig. 7). It is nevertheless possible that both coregulation and interdigitation act on several of the loci we have presented here, especially those duplicated GRBs that have retained more than one gene in both copies after teleost whole-genome duplication.

GRBs and disease breakpoints

Since developmental regulatory genes are part of GRBs that need to be kept intact to maintain correct gene expression, we asked whether GRBs can be used to search for the likely target genes of position effect disease mutations in the human genome. A recent study (Ahituv et al. 2005) demonstrated the utility of using conserved blocks of synteny to establish likely genomic ranges in which to look for particular position effect mutations. We have described experimental indications that insertions in zebrafish can be used to study those mutations, and that the computational analysis can help locate their likely target genes.

To date, very few human position effect mutations have been identified, among them aniridia, a chromosomal rearrangement downstream from human *PAX6* and inside *ELP4* (Kleinjan et al. 2001). In the zebrafish, an insertion inside *elp4* takes on the expression pattern of *pax6.2*, showing that this gene is inside the regulatory domain of *pax6.2*. Likewise, an insertion inside the *fgf8* bystander gene *fbxw4* takes on the expression pattern of *fgf8*. Transposon insertions in *Fbxw4* were determined to be causal in the mouse semidominant *dactylaplasia* mutation, in the absence of mutations in the coding sequence (Sidow et al. 1999). *Fgf8* expression in the apical ectodermal ridge (AER) is not properly maintained in the mouse mutants, and this expression defect correlates well with the observed phenotype in both mouse *dactylaplasia* and the corresponding human genetic disease split hand/foot malformation 3 (OMIM#600095), which maps to the *FGF8* GRB in the human genome (de Mollerat et al. 2003). *Fgf8* has been shown in the mouse to be the only *Fgf* family member expressed in the AER and necessary for normal limb development (Lewandoski et al. 2000). Intronic insertions in *fbxw4* in zebrafish cause a semidominant adult pigment stripe pattern defect (Kawakami et al. 2002), which we also found with insertions CLGY1030, CLGY508, and CLGY667 (Fig. 3A; data not shown). Thus, although the phenotype in zebrafish is different from in mouse, these data suggest that *fbxw4* is a bystander gene and that the defect underlying these mutations is misregulation of *fgf8*. It is therefore reasonable to speculate that *FBXW4* has been incorrectly assigned as the disease gene in human split hand/foot malformation 3.

The search for putative regulatory elements or previously unknown exons can generate target sequences to be resequenced in patient DNA. In one of the two cases that were characterized bioinformatically, *otp*, we searched for mapped human diseases at the edge of the *otp* GRB and found Hermansky-Pudlak Syndrome, type 2, a cell migration and platelet defect (OMIM#608233) mapped to the *AP3B1* gene at 5q14.1, which, as we have shown, contains HCNEs of the *otp* GRB. Recently, a microdeletion of ~8 kb causing this disease was mapped in exons 14–15 of *AP3B1* (Jung et al. 2006), which removes one of the HCNEs mapped in

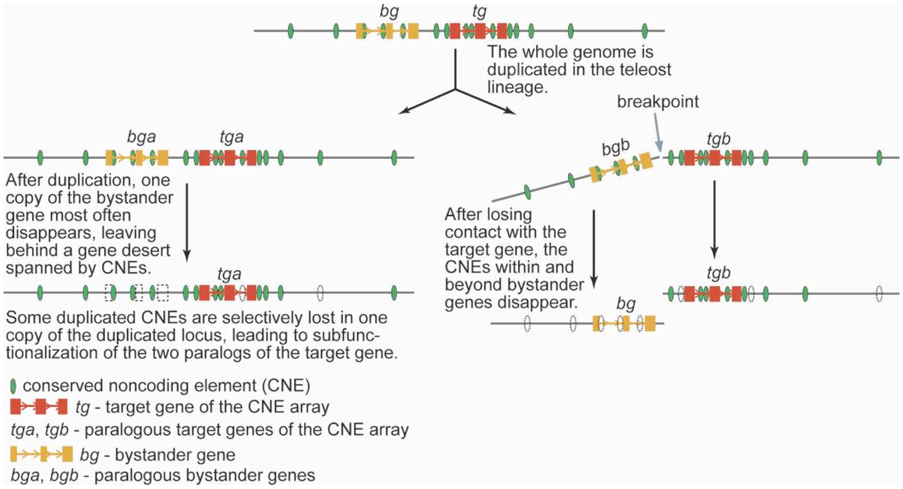
bg
tg
The whole genome is duplicated in the teleost lineage.
breakpoint
bga
tga
bgb
tgb
After duplication, one copy of the bystander gene most often disappears, leaving behind a gene desert spanned by CNEs.
After losing contact with the target gene, the CNEs within and beyond bystander genes disappear.
tga
tgb
bg
Some duplicated CNEs are selectively lost in one copy of the duplicated locus, leading to subfunctionalization of the two paralogs of the target gene.
conserved noncoding element (CNE)
tg - target gene of the CNE array
tga, tgb - paralogous target genes of the CNE array
bg - bystander gene
bga, bgb - paralogous bystander genes

Figure 7. A possible scenario of teleost GRB duplication conceptually corresponds to the case of the target gene *OTP* and its bystander genes *TBCA* and *AP3B1* (Fig. 2A). In one case (*left*), the bystander gene (bg) disappears by neutral evolution. This scenario directly demonstrates that HCNEs within and beyond the bystander gene do not control that gene but the target gene. Other scenarios are variations on the theme—for instance, after breaking off one of the copies of a bystander gene (or a set of bystander genes) from the corresponding copy of the target gene (*right*), it can be the broken-off copy that disappears, together with the HCNEs that lost contact with the target gene.

this study. While the *AP3B1* mutation may have an effect on the disease phenotype, this finding suggests that the disease may be a compound phenotype of loss of function of *AP3B1* and loss of specific regulatory input of *OTP*. Thus, as demonstrated by this case, establishment of human/teleost conserved synteny combined with fine mapping in the human genome can immediately produce strong candidate targets for human position effect mutations. This is in accord with observations by Ahituv et al. (2005), but additionally facilitates identification of likely target genes affected by position effects among unaffected bystander genes.

Mutations affecting expression of microRNAs in genetic disease have so far not been reported. However, since they can be target genes in GRBs, such cases may exist. In the case of the two GRBs described here that contain elements driving the transcriptional activity of microRNAs of the *mirn9* family, there are human disease loci mapped to the area. In the case of *mirn9-2*, there is a cone rod dystrophy mapped to 1q12–1q24 (OMIM%605549), where so far no gene has been assigned. Close to *mirn9-5* is Usher syndrome type II and febrile seizures at 5q14 (OMIM#605472 and #604352). Some cases of these diseases have been assigned to mutations in the gene encoding G-coupled receptor MASS1 (for example, see Nakayama et al. 2002), which lies within the larger synteny block containing *mirn9-5* and is ~1.8 Mb upstream of *MEF2C*, but other cases remain unresolved. Expression of the *mirn9* family is observed throughout the CNS and retina in zebrafish (Wienholds et al. 2005). It is interesting that both of these miRNAs are also near myocyte enhancer factor genes, which are also thought to be of developmental regulatory function. Why the miRNAs and the *mefs* are kept together is currently not known, but it may be that they are regulated together, as mouse *Mef2c* and *Mef2d* are expressed in the telencephalon (Lyons et al. 1995). In the case of the gene desert containing zebrafish *id1*, annotation to the human genome reveals an orthologous relationship between zebrafish *id1* and human *ID1*, even though their similarity at the protein level is low. OMIM lists an ataxia in the area of *ID1* in the human genome at 20q11 (OMIM%608029) (Tranebjaerg et al. 2003), and *id1* is expressed in the developing cerebellum. For *rx3*, the human ortholog *RAX* is embedded in a large syntenic block, and a cone-rod dystrophy (OMIM%600624) has been mapped to the area (Warburg et al. 1991).

In the *BARHL1* GRB, the far upstream *TSC1* gene at 9q34 is implicated in tuberous sclerosis (OMIM#191100) and focal cortical dysplasia of Taylor (OMIM#607341), distinguished by epileptic seizures and likely caused by a neuronal migration defect (Wolf et al. 1995), which is consistent with the *barhl1* expression pattern. However, the *TSC1* ortholog was duplicated in teleosts along with *barhl1* and might be a developmental regulator itself, perhaps coregulated with *barhl1* orthologs. The data presented in this paper suggest that for the mapping of human diseases it will be important to establish whether the implied disease gene is a GRB target gene or is, in fact, located in a GRB as a bystander with no functional relation to the regulatory inputs of the enhancers of the GRB. In such cases, it will be important to correlate GRB regulatory content with the disease phenotype and, if warranted, reassign the disease phenotype to the correct gene. Thus, important developmental genes are embedded in large GRBs, breakpoints or mutations within these GRBs may cause genetic disease, and subsequent fine mapping may result in the indictment of a bystander gene containing essential regulatory elements for a distant target gene.

The bystander genes, although within reach of specific regulatory elements, are expressed in different patterns and thus are not specifically regulated by GRB regulatory content. How this apparent specificity comes about is currently not understood.

Synteny and vertebrate genome evolution

Nadeau and Taylor (1984) suggested a model of genome evolution in which evolutionary chromosomal breakpoints are distributed randomly throughout the mouse and human genomes, and postulated conserved blocks of synteny to be "relics of ancient linkage groups not yet disrupted by chromosome rearrangement." In closely related genomes such as mouse and human, this may be partially true, but it is likely that in distantly related vertebrates, sufficient numbers of translocation events have occurred during evolution to rearrange all large chromosomal regions. The Nadeau and Taylor paper was published before the discovery of *hox* clusters and very large genes, neither of which can be broken without disease as a result, and thus are exceptions to this hypothesis. It was recently noted that there are synteny blocks of significant size across all vertebrate genomes, and these have been hypothesized to result from the need to be kept intact by regulatory sequences (MacKenzie et al. 2004; Ahituv et al. 2005; Goode et al. 2005; Kleinjan and van Heyningen 2005; Gomez-Skarmeta et al. 2006). These and the results in this study suggest that the Nadeau and Taylor hypothesis is not plausible for the explanation of synteny in general.

The comparative analysis of zebrafish gene maps indicated (Postlethwait et al. 2000; Woods et al. 2000) and the subsequent sequencing of the *Tetraodon* genome (Jaillon et al. 2004) confirmed an ancient whole-genome duplication event, followed by loss of most of the duplicated genes. However, for the synteny comparisons between tetraodon and human genomes, these authors examined gene order, not underlying noncoding sequence similarity. It is important to note that synteny, as we have shown here, while typically defined as conserved gene linkage, is often rather the conservation of order of the underlying regulatory elements. The general rule for the destiny of GRBs after genome duplication seems to involve loss of individual HCNEs in one copy of the GRB (Fig. 7). When an entire part of a HCNE array is detached from a target gene by chromosomal rearrangement, neighboring ubiquitously expressed genes that harbor HCNEs in their noncoding sequence will be retained in the detached segment, and the HCNEs will be lost from that segment. Any alternative explanations that would account for the observed disentangling of regulatory and protein-coding elements are highly improbable.

Recently, early developmental regulators were found to be associated with transposon-free regions (TFRs) in the human genome; for instance, *PAX6* and *MIRN9-2* (Simons et al. 2006). However, we do not find such a correlation: we found two retroviral insertions within the third intron of *pax6.2*, with no detectable phenotype. On the other hand, the area around *fgf8*, which we found not to tolerate insertions in zebrafish (data not shown), contains evidence of numerous transposons in both human and fish genomes. Thus, while it is intriguing that developmental regulatory genes are associated with TFRs in the human genome, their implication in long-range regulation is not straightforward and may be spacing- and site-dependent.

Position effects were first demonstrated in *Drosophila*, where tandem duplications at the *barh* locus cause a dominant eye defect (Sturtevant 1925). Remarkably, *barh* is an ortholog of human

BARHL1 (and *BARHL2*). We propose that position effect mutations in *Drosophila* as well as in vertebrates are disturbances of GRBs. Considering that synteny is a feature present across *Drosophila* genomes, we postulate that GRBs will also be found in Drosophilids, the species where chromosomal gene order was first demonstrated (Sturtevant 1913).

The above examples demonstrate that genomic regulatory blocks play an essential role not only in the regulation of activity of developmental genes, but also in the evolutionary dynamics of entire chromosomal loci by imposing long-range constraints on their structure and integrity. Whole-genome duplication can transiently relieve those constraints and enable neighboring genes to "escape" the gridlock imposed by long-range regulatory elements. Teleost genomes provide a fertile ground for studying this phenomenon in detail.

Methods

Enhancer detection

Viral insertions into the zebrafish germline, screening, and identification of chromosomal insertion sites were done as described (Ellingsen et al. 2005; Laplante et al. 2006). The insertions described in this study were generated in a large-scale screen, which examined ~10,000 random insertions, of which ~1500 were active, ~900 were kept, and 350 were mapped. For the insertions in this study, all transgenic lines with similar expression patterns were selected and the insertion sites identified. The flanking sequences of all insertions are listed in Supplemental Table S3. All experiments were in accordance with regulations for animal experimentation in Norway.

Expression patterns

Expression data for this paper were retrieved from the Zebrafish Information Network (ZFIN), the Zebrafish International Resource Center (University of Oregon, Eugene; http://zfin.org/), during the course of this study. Additional in situ hybridizations were done as described (Thisse and Thisse 1998).

Sequence and annotation data

We used the following genome assemblies: human genome build NCBI 36.1, zebrafish genome build Zv6, *Tetraodon* genome build V7, and chicken genome build V1. Ensembl genes, miRNA and OMIM, and Gene Ontology (GO) annotation were obtained from Ensembl 39 (Birney et al. 2006), and net alignments and remaining genome annotations were obtained from the UCSC Genome Browser database (Karolchik et al. 2003). The zebrafish genome sequence and gene annotation were produced by the Wellcome Trust Sanger Institute. The data can be accessed through the Sanger Institute Web resources (http://www.sanger.ac.uk/Projects/D_rerio/). The annotation followed the procedures described in Jekosch (2004). Annotated contigs were accessed and aligned in the Ensembl (Birney et al. 2006) and Vega (Ashurst et al. 2005) databases.

HCNE detection

We identified HCNEs conserved between human and zebrafish by scanning a zebrafish-to-human net alignment (Kent et al. 2003) for maximal regions with at least 70% sequence identity and a minimum length of 50 bp. Human-to-chicken HCNEs were likewise extracted using a 90% sequence identity threshold. We discarded elements whose human genome coordinates overlapped by one or more base pairs with any exon in Ensembl 39 protein-coding genes, RefSeq genes, UCSC known genes, or GENSCAN predictions. The remaining conserved elements were considered HCNEs. We similarly produced two sets of zebrafish–tetraodon HCNEs by scanning a zebrafish-to-tetraodon net alignment, using sequence identity thresholds of 70% and 90%, respectively, and excluded those that overlapped exons in any of the following zebrafish genome annotations: Ensembl 39 protein-coding genes, RefSeq genes, GENSCAN predictions, zebrafish mRNAs or ESTs from RefSeq or GenBank, non-zebrafish mRNAs from GenBank, or human proteins. HCNEs in this paper are listed in Supplemental Table S2.

Detection of synteny blocks between the zebrafish and human genomes

Previous approaches to detect synteny blocks between human and fish were based on transcript or protein sequence comparisons (Aparicio et al. 2002; Jaillon et al. 2004; Woods et al. 2005). Because we were interested in rearrangements of both coding and noncoding sequence, we wished to define synteny based on direct genome sequence comparisons. Such approaches have been described, for example, for human–mouse synteny (Waterston et al. 2002; Pevzner and Tesler 2003), and have typically been based on high-scoring reciprocal-best alignments between genomes. Reciprocal-best alignments, however, are not ideal for human–teleost comparisons because of the whole-genome duplication that has occurred in the teleost lineage (Jaillon et al. 2004; Woods et al. 2005). We therefore based our synteny blocks on net alignments (Kent et al. 2003) from the zebrafish genome to the human genome. Since neutrally evolving sequence typically cannot be aligned between human and zebrafish genomes, many syntenic regions are divided over several alignments separated by large regions of unaligned sequence. This segregation is augmented by the presence of many local assembly errors in the zebrafish genome assembly. The net alignment procedure allows gaps to some degree, but to allow for inversions and other local rearrangements such that syntenic blocks are separated by macrorearrangements rather than smaller insertions and alignment gaps, we constructed a graph based on the highest-scoring (level 1) net alignments where two alignments (nodes) were connected if they were separated by <100 kb in the zebrafish genome and <300 kb in the human genome. We then considered each connected component in the graph to be one synteny block. We kept the synteny block with most aligned bases to the human genome in cases of block overlap in the zebrafish genome. We did not set a lower bound on synteny block size, but accounted for the genomic span of synteny blocks in all downstream analyses. Our conclusions are not dependent on these particular threshold settings, but can be reconfirmed using a range of thresholds (data not shown).

Analysis of synteny for different biological processes

For each protein-coding zebrafish gene in Ensembl, we computed the extent of synteny around it, defined as the genomic span of the synteny block in which the gene is contained, excluding the region spanned by the gene itself (to control for differences in gene size). The category "any biological process" contains all genes annotated with a GO biological process term other than "biological process unknown." The *hox* and *irx* families of developmental regulatory genes were excluded from the analysis because they are known to be kept together in large synteny blocks to maintain coregulation. We assigned a gene to a synteny block if that gene had one transcript with at least 95% of its coding sequence spanned by the synteny block and at least 50% of its coding sequence aligned, at the resolution of the net alignment track, to the syntenic locus in the human genome.

Acknowledgments

We thank D. Chourrout, U. Technau, E. Thompson, J.L. Gómez-Skarmeta, and S. Rinkwitz for helpful comments on the manuscript. We also thank Eilen Myrvold, Merete Nilsen, Heikki Savolainen, and Tarjei Raniseth in the Sars Centre fish facilities for expert technical support. This work was supported by funds from the Sars Centre, the University of Bergen, and by the Functional Genomics Program (FUGE) in the Research Council of Norway to T.S.B. and B.L. and by the European Commission as part of the ZF-Models integrated project in the 6th Framework Programme (Contract No. LSHG-CT-2003-503469) to T.S.B., M.C., I.S., K.H., C.T., and B.T.; and by EMBO short-term fellowships to J.G. and A.G.; an HFSP short-term fellowship to P.M.; and a FEBS short-term fellowship to B.A. The zebrafish genome sequencing project at the Sanger Institute is funded by the Wellcome Trust.

References

Acampora, D., Postiglione, M.P., Avantaggiato, V., Di Bonito, M., Vaccarino, F.M., Michaud, J., and Simeone, A. 1999. Progressive impairment of developing neuroendocrine cell lineages in the hypothalamus of mice lacking the *Orthopedia* gene. *Genes & Dev.* **13:** 2787–2800.

Ahituv, N., Prabhakar, S., Poulin, F., Rubin, E.M., and Couronne, O. 2005. Mapping *cis*-regulatory domains in the human genome using multi-species conservation of synteny. *Hum. Mol. Genet.* **14:** 3057–3063.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **299:** 1301–1310.

Ashurst, J.L., Chen, C.K., Gilbert, J.G., Jekosch, K., Keenan, S., Meidl, P., Searle, S.M., Stalker, J., Storey, R., Trevanion, S., et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* **33:** D459–D465.

Bellen, J.H. 1999. Ten years of enhancer detection: Lessons from the fly. *Plant Cell* **11:** 2271–2281.

Bellen, H.J., Levis, R.W., Liao, G., He, Y., Carlson, J.W., Tsang, G., Evans-Holm, M., Hiesinger, P.R., Schulze, K.L., Rubin, G.M., et al. 2004. The BDGP gene disruption project: Single transposon insertions associated with 40% of *Drosophila* genes. *Genetics* **167:** 761–781.

Biemar, F., Zinzen, R., Ronshaugen, M., Sementchenko, V., Manak, J.R., and Levine, M.S. 2005. Spatial regulation of microRNA gene expression in the *Drosophila* embryo. *Proc. Natl. Acad. Sci.* **102:** 15907–15911.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., et al. 2006. Ensembl 2006. *Nucleic Acids Res.* **34:** D556–D561.

Boffelli, D., Nobrega, M.A., and Rubin, E.M. 2004. Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* **5:** 456–465.

Bulfone, A., Menguzzato, E., Broccoli, V., Marchitiello, A., Gattuso, C., Mariani, M., Consalez, G.G., Martinez, S., Ballabio, A., and Banfi, S. 2000. *Barhl1*, a gene belonging to a new subfamily of mammalian homeobox genes, is expressed in migrating neurons of the CNS. *Hum. Mol. Genet.* **22:** 1443–1452.

Carvajal, J.J., Cox, D., Summerbell, D., and Rigby, P.W. 2001. A BAC transgenic analysis of the Mrf4/Myf5 locus reveals interdigitated elements that control activation and maintenance of gene expression during muscle development. *Development* **128:** 1857–1868.

Colombo, A., Reig, G., Mione, M., and Concha, M.L. 2006. Zebrafish BarH-like genes define discrete neural domains in the early embryo. *Brain Res. Gene Expr. Patterns* **6:** 347–352.

Conaco, C., Otto, S., Han, J.J., and Mandel, G. 2006. Reciprocal actions of REST and a microRNA promote neuronal identity. *Proc. Natl. Acad. Sci.* **103:** 2422–2427.

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.

de Mollerat, X.J., Gurrieri, F., Morgan, C.T., Sangiorgi, E., Everman, D.B., Gaspari, P., Amiel, J., Bamshad, M.J., Lyle, R., Blouin, J.L., et al. 2003. A genomic rearrangement resulting in a tandem duplication is associated with split hand-split foot malformation 3 (SHFM3) at 10q24. *Hum. Mol. Genet.* **12:** 1959–1971.

Ellingsen, S., Laplante, M.A., Konig, M., Kikuta, H., Furmanek, T., Hoivik, E.A., and Becker, T.S. 2005. Large-scale enhancer detection in the zebrafish genome. *Development* **132:** 3799–3811.

Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., and McCallion, A.S. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312:** 276–279.

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.

Glaser, T., Walton, D.S., and Maas, R.L. 1992. Genomic structure, evolutionary conservation and aniridia mutations in the human PAX6 gene. *Nat. Genet.* **2:** 232–239.

Gomez-Skarmeta, J.L., Lenhard, B., and Becker, T.S. 2006. New technologies, new findings, and new concepts in the study of vertebrate *cis*-regulatory sequences. *Dev. Dyn.* **235:** 870–885.

Goode, D.K., Snell, P., Smith, S.F., Cooke, J.E., and Elgar, G. 2005. Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics* **86:** 172–181.

Inoue, F., Nagayoshi, S., Ota, S., Islam, M.E., Tonou-Fujimori, N., Odaira, Y., Kawakami, K., and Yamasu, K. 2006. Genomic organization, alternative splicing, and multiple regulatory regions of the zebrafish fgf8 gene. *Dev. Growth Differ.* **48:** 447–462.

Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431:** 946–957.

Jekosch, K. 2004. The zebrafish genome project: Sequence analysis and annotation. *Methods Cell Biol.* **77:** 225–239.

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

Jordan, T., Hanson, I., Zaletayev, D., Hodgson, S., Prosser, J., Seawright, A., Hastie, N., and van Heyningen, V. 1992. The human PAX6 gene is mutated in two patients with aniridia. *Nat. Genet.* **1:** 328–332.

Jung, J., Bohn, G., Allroth, A., Boztug, K., Brandes, G., Sandrock, I., Schaffer, A.A., Rathinam, C., Kollner, I., Beger, C., et al. 2006. Identification of a homozygous deletion in the AP3B1 gene causing Hermansky Pudlak syndrome, type 2. *Blood* **108:** 362–369.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser database. *Nucleic Acids Res.* **31:** 51–54.

Kawakami, K., Amsterdam, A., Shimoda, N., Becker, T., Mugg, J., Shima, A., and Hopkins, N. 2002. Proviral insertions in the zebrafish *hagoromo* gene, encoding an F-box/WD40-repeat protein, cause stripe pattern anomalies. *Curr. Biol.* **10:** 463–466.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484–11489.

Kleinjan, D.A. and van Heyningen, V. 2005. Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76:** 8–32.

Kleinjan, D.A., Seawright, A., Schedi, A., Quinlan, R.A., Danes, S., and van Heyningen, V. 2001. Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Hum. Mol. Genet.* **10:** 2049–2059.

Laplante, M., Kikuta, H., Konig, M., and Becker, T.S. 2006. Enhancer detection in the zebrafish using pseudotyped murine retroviruses. *Methods* **39:** 189–198.

Lee, A.P., Koh, E.G., Tay, A., Brenner, S., and Venkatesh, B. 2006. Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. *Proc. Natl. Acad. Sci.* **103:** 6994–6999.

Lewandoski, M., Sun, X., and Martin, G.R. 2000. Fgf8 signalling from the AER is essential for normal limb development. *Nat. Genet.* **26:** 460–463.

Loots, G.G., Kneissel, M., Keller, H., Baptist, M., Chang, J., Collette, N.M., Ovcharenko, D., Plajzer-Frick, I., and Rubin, E.M. 2005. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res.* **15:** 928–935.

Lyons, G.E., Micales, B.K., Schwarz, J., Martin, J.F., and Olson, E.N. 1995. Expression of mef2 genes in the mouse central nervous system suggests a role in neuronal maturation. *J. Neurosci.* **15:** 5727–5738.

MacKenzie, A., Miller, K.A., and Collinson, J.M. 2004. Is there a functional link between gene interdigitation and multi-species conservation of synteny blocks? *Bioessays* **26:** 1217–1224.

Mathers, P.H., Grinberg, A., Mahon, K.A., and Jamrich, M. 1997. The Rx homeobox gene is essential for vertebrate eye development. *Nature* **387:** 603–607.

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

Nadeau, J.H. and Taylor, B.A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci.* **81:** 814–818.

Nakayama, J., Fu, Y.-H., Clark, A.M., Nakahara, S., Hamano, K., Iwasaki, N., Matsui, A., Arinami, T., and Ptacek, L.J. 2002. A nonsense mutation of the MASS1 gene in a family with febrile and afebrile seizures. *Ann. Neurol.* **52:** 654–657.

Ohno, S. 1973. Ancient linkage groups and frozen accidents. *Nature* **244:** 259–262.

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15:** 137–145.

Peng, Q., Pevzner, P.A., and Tesler, G. 2006. The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comput. Biol.* **2:** e14.

Pevzner, P. and Tesler, G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci.* **100:** 7672–7677.

Postlethwait, J.H., Woods, I.G., Ngo-Hazelett, P., Yan, Y.L., Kelly, P.D., Chu, F., Huang, H., Hill-Force, A., and Talbot, W.S. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* **10:** 1890–1902.

Reifers, F., Adams, J., Mason, I.J., Schulte-Merker, S., and Brand, M. 2000. Overlapping and distinct functions provided by fgf17, a new zebrafish member of the Fgf8/17/18 subgroup. *Mech. Dev.* **99:** 39–49.

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.

Sankoff, D. and Trinh, P. 2005. Chromosomal breakpoint reuse in genome sequence rearrangement. *J. Comput. Biol.* **12:** 812–821.

Sidow, A., Bulotsky, M.S., Kerrebrock, A.W., Birren, B.W., Altshuler, D., Jaenisch, R., Johnson, K.R., and Lander, E.S. 1999. A novel member of the F-box/WD40 gene family, encoding dactylin, is disrupted in the mouse dactylaplasia mutant. *Nat. Genet.* **23:** 104–107.

Simons, C., Pheasant, M., Makunin, I.V., and Mattick, J.S. 2006. Transposon-free regions in mammalian genomes. *Genome Res.* **16:** 164–172.

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

Stigloher, C., Ninkovic, J., Laplante, M., Geling, A., Tannhauser, B., Topp, S., Kikuta, H., Becker, T.S., Houart, C., and Bally-Cuif, L. 2006. Segregation of telencephalic and eye-field identities inside the zebrafish forebrain territory is controlled by Rx3. *Development* **133:** 2925–2935.

Stone, E.M., Lotery, A.J., Munier, F.L., Heon, E., Piguet, B., Guymer, R.H., Vandenburgh, K., Cousin, P., Nishimura, D., Swiderski, R.E., et al. 1999. A single EFEMP1 mutation associated with both Malattia Leventinese and Doyne honeycomb retinal dystrophy. *Nat. Genet.* **22:** 199–202.

Sturtevant, A.H. 1913. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J. Exp. Zool.* **14:** 43–59.

Sturtevant, A.H. 1925. The effects of unequal crossing over at the bar locus in *Drosophila*. *Genetics* **10:** 117–147.

Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C., Ma, H., and Martienssen, R. 1995. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes & Dev.* **9:** 1797–1810.

Thisse, C. and Thisse, B. 1998. High-resolution whole-mount in situ hybridization. In *Zebrafish Science Monitor*, Vol. 5. University of Oregon Press, Eugene. http://zfin.org/zf_info/monitor/vol5.1/vol5.1.html.

Tranebjaerg, L., Teslovich, T.M., Jones, M., Barmada, M.M., Fagerheim, T., Dahl, A., Escolar, D.M., Trent, J.M., Gillanders, E.M., and Stephan, D.A. 2003. Genome-wide homozygosity mapping localizes a gene for autosomal recessive non-progressive infantile ataxia to 20q11–q13. *Hum. Genet.* **113:** 293–295.

Vavouri, T., McEwen, G.K., Woolfe, A., Gilks, W.R., and Elgar, G. 2006. Defining a genomic radius for long-range enhancer action: Duplicated conserved non-coding elements hold the key. *Trends Genet.* **22:** 5–10.

Voronina, V.A., Kozhemyakina, E.A., O'Kernick, C.M., Kahn, N.D., Wenger, S.L., Linberg, J.V., Schneider, A.S., and Mathers, P.H. 2004. Mutations in the human RAX homeobox gene in a patient with anophthalmia and sclerocornea. *Hum. Mol. Genet.* **13:** 315–322.

Warburg, M., Sjo, O., and Tranebjaerg, L. 1991. Chromosome 6q deletion and retinal cone dystrophy. *Am. J. Med. Genet.* **38:** 134.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexanersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Wienholds, E., Kloosterman, W.P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H.R., Kauppinen, S., and Plasterk, R.H. 2005. MicroRNA expression in zebrafish embryonic development. *Science* **309:** 310–311.

Wolf, H.K., Wellmer, J., Muller, M.B., Wiestler, O.D., Hufnagel, A., and Pietsch, T. 1995. Glioneuronal malformative lesions and dysembryoplastic neuroepithelial tumors in patients with chronic pharmacoresistant epilepsies. *J. Neuropathol. Exp. Neurol.* **54:** 245–254.

Woods, I.G., Kelly, P.D., Chu, F., Ngo-Hazelett, P., Yan, Y.L., Huang, H., Postlethwait, J.H., and Talbot, W.S. 2000. A comparative map of the zebrafish genome. *Genome Res.* **10:** 1903–1914.

Woods, I.G., Wilson, C., Friedlander, B., Chang, P., Reyes, D.K., Nix, R., Kelly, P.D., Chu, F., Postlethwait, J.H., and Talbot, W.S. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res.* **15:** 1307–1314.

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.

Zerucha, T. and Ekker, M. 2000. Distal-less-related homeobox genes of vertebrates: evolution, function, and regulation. *Biochem. Cell Biol.* **78:** 593–601.

Received October 31, 2006; accepted in revised form January 22, 2007.