Large-scale enhancer detection in the zebrafish genome

Staale Ellingsen*, Mary A. Laplante, Melanie König, Hiroshi Kikuta, Tomasz Furmanek, Erling A. Hoivik† and Thomas S. Becker‡

Sars International Centre for Marine Molecular Biology at the University of Bergen, Thormoehlensgate 55, 5008 Bergen, Norway
*Present address: EMBL, Developmental Biology Programme, Meyerhofstrasse1, 69117 Heidelberg, Germany
†Present address: Institute for Biomedicine, University of Bergen, Jonas Lies Vei 91, 5008 Bergen, Norway
‡Author for correspondence (e-mail: tom.becker@sars.uib.no)

Accepted 29 June 2005

Development 132, 3799-3811
Published by The Company of Biologists 2005
doi:10.1242/dev.01951

Summary

Murine retroviral vectors carrying an enhancer detection cassette were used to generate 95 transgenic lines of fish in which reporter expression is observed in distinct patterns during embryonic development. We mapped 65 insertion sites to the as yet unfinished zebrafish genome sequence. Many integrations map close to previously known developmental genes, including transcription factors of the Pax, Hox, Sox, Pou, Otx, Emx, zinc-finger and bHLH gene families. In most cases, the activated provirus is located in, or within a 15 kb interval around, the corresponding transcriptional unit. The exceptions include four insertions into a gene desert on chromosome 20 upstream of *sox11b*, and an insertion upstream of *otx1*. In these cases, the activated insertions are found at a distance of between 32 kb and 132 kb from the coding region. These as well as seven other insertions described here identify genes that have recently been associated with ultra conserved non-coding elements found in all vertebrate genomes.

Key words: Cis-regulatory sequence, Synteny, In vivo imaging, Zebrafish

Introduction

The recent analysis of the human and several other vertebrate genomes has brought with it significant advances in the identification of evolutionarily conserved, and therefore functional, sequences. For example, comparison with the mouse sequence suggests that about 3.3% of the human genome encode exons (Dermitzakis et al., 2005), and intense efforts are under way at multiple levels to annotate, on a genomic scale, the function and expression profiles of all vertebrate genes. It is an established idea that tissue-specific gene transcription is driven by cis-regulatory sequences termed enhancers, and hence that gene expression can be defined experimentally, to a first approximation, through the activity of these elements. The classic example of such work is enhancer detection in *Drosophila*, a pioneering methodology using a randomly inserting transposon-derived vector encoding a reporter protein downstream of a minimal promoter (Bellen, 1999; Bier et al., 1989; Wilson et al., 1989; O'Kane and Gehring, 1987). Upon insertion of the vector into the fly genome within operating distance of *cis*-regulatory elements, expression of the reporter can be observed in transgenic animals, giving insight into nearby genes that are regulated by these sequences (Bellen, 1999; Bier et al., 1989; Wilson et al., 1989; O'Kane and Gehring, 1987; Spradling et al., 1999). This technique has greatly enhanced the discovery and characterization of cell types, and the genes involved in their determination. Consequently, *Drosophila* cis-regulation, as well as the associated genes, is among the best understood of all metazoans, resulting in the analysis of entire gene regulation networks (e.g. Schroeder et al., 2004).

Spurred by these successes, efforts have been aimed at establishing enhancer detection in vertebrate model organisms (Korn et al., 1992; Bayer et al., 1992). More recently, transposon based enhancer detection protocols have been reported for Medaka and for the zebrafish (Grabher et al., 2003; Balciunas et al., 2004; Parinov et al., 2004), but to date very few insertion sites have been characterized, and no large-scale effort has been attempted. Engineered murine leukemia retroviruses (MLV) represent the most efficient insertional agents in vertebrate systems to date, and have been developed as gene delivery vectors for gene therapy, insertional mutagenesis and other experimental approaches (Frankel et al., 1985; Jahner et al., 1982; Sanes, 1989; Austin and Cepko, 1994; Gaiano et al., 1996a; Gaiano et al., 1996b; Pfeifer and Verma, 2001; Amsterdam et al., 2004). When pseudotyped with the Vesicular Stomatitis Virus G protein (VSV-G), MLV can infect zebrafish cells (Gaiano et al., 1996a; Burns et al., 1993; Lin et al., 1994) and expression from an internal ubiquitous promoter was shown to be detectable after integration and subsequent germline passage (Linney et al., 1999). This latter finding suggested that MLV proviruses are not transcriptionally silenced in the zebrafish genome, in contrast to what has been observed in their normal host, the mouse (Jahner and Janisch, 1985). We have devised and report here an MLV-derived enhancer detection vector containing an internal, basal zebrafish promoter upstream of a yellow fluorescent protein (YFP) reporter gene. We demonstrate that this vector expresses the reporter gene in a subset of genomic integrations in cultured zebrafish cells, as well as in zebrafish embryos derived from parents carrying proviral germline insertions. A recent genetic screen in our laboratory has

generated around 1000 transgenic lines of zebrafish that express fluorescent proteins in early tissue specific patterns, and we show here that the corresponding insertions can be mapped onto the unfinished zebrafish genome. Many of the insertions have occurred close to developmental regulatory genes, exemplified by a number of known transcriptional regulators. The method and the results described in this paper represent the first approach to experimentally characterize regions of any vertebrate genome with cis-regulatory activity and the genes therein on a genomic scale and will probably enhance our understanding of transcriptional regulation during vertebrate, including human, embryonic development.

Materials and methods

Construction of viral plasmid

The enhancer detection vector is based on a Murine Leukemia Virus (MLV) type retroviral vector (pCL, gift of Dr Inder Verma) (Naviaux et al., 1996). To construct the enhancer detection vector pCLGY, the gene for yellow fluorescent protein (YFP) was ligated into the *Bam*HI site of the empty vector. Second, a 1 kb proximal promoter of the zebrafish GATA2 gene (Meng et al., 1997) was ligated into the vector upstream of the reporter sequence. The transcriptional direction of the inserted basal promoter-reporter cassette is the same as that of the virus.

Production of VSV-G pseudotyped retrovirus

Generation of pseudotyped viruses was carried out as previously described (Chen et al., 2002). The viral vector pCLGY and a construct expressing the envelope protein VSV-G (Burns et al., 1993) were co-transfected into a 293 gag-pol packaging cell line (293 gp/bsr, gift of Dr Inder Verma). Virus was harvested 48 hours post transfection, and concentrated by ultracentrifugation at 50,000 ***g*** at 4°C for 90 minutes.

Analysis of zebrafish cells infected with retrovirus

Cultured zebrafish Pac2 fibroblast cells (Chen et al., 2002) were infected with the VSV-G pseudotyped CLGY virus, and analyzed for fluorescence 48 hours post infection by flow cytometry, using a FACScalibur analyzer (Becton Dickinson, USA). The numbers of integrated virus per cell were calculated using real-time PCR, essentially as previously described (Chen et al., 2002). In brief, genomic DNA was isolated from the same batch of infected cells as used for flow cytometry. Proviral sequence was amplified using previously described primers and probes (Amsterdam et al., 1999). To normalize for the DNA amount used as template for the PCR reactions, an endogenous sequence was co-amplified with the proviral sequence, using the following primers and probes: forward, GTATGCCAACAAAGGCAGCA; reverse TGGGTTTTCTGGTTC-CAGGT; Taqman probe, Yakima Yellow-CCCATCGAGCAGATC-CCCGA-Darquencher.

Generation and identification of transgenic founder fish and F1 embryos

Zebrafish were obtained from our breeding colony kept and raised according to a protocol developed in our laboratory (www.sars.no/manual.doc). Zebrafish embryos were dechorionated at early blastula stages in 1× Holtfreters solution containing pronaseE (Sigma). Ten to 20 nl of the concentrated virus, containing 4 μg of polybrene per ml, were injected at three or four locations among the cells of blastula-stage zebrafish embryos (~500-2000 cell stage), as previously described (Gaiano et al., 1996a; Chen et al., 2002). After injection, embryos were incubated at 37°C for 2-4 hours before being transferred to 28.5°C for further raising.

Injected founders (F0 generation) were raised to sexual maturity and outcrossed to wild-type fish. F1 larvae were screened at 24 hours post fertilization using a TE2000-S inverted microscope (Nikon) equipped with 10× and 20× lenses, and a 500/20 nm excitation filter and a 515 nm BP emission filter (Chroma) for detection of YFP. Photographs of live positive embryos were taken using a Spot monochrome digital camera and associated software (Diagnostic Systems). Images were processed in Adobe Photoshop by adjusting levels. High resolution images of the lines in this paper are available at http://clgy.no/clgyimages/.

Cloning of flanking sequences from activated integrated vectors

Eight YFP-positive embryos from each enhancer detection line were raised until 5 dpf. Genomic DNA was isolated from each individual larva and flanking genomic sequence of the activated viral integration was amplified using linker mediated PCR (LM-PCR) (Wu et al., 2003). Genomic DNA was digested with *Mse*I and the resulting fragments were ligated to an *Mse*I linker. LM-PCR was performed with one primer specific to the viral LTR and the other primer specific to the linker. Nested PCR was then performed using a second pair of primers internal to the first primer pair. Based on agarose gel analysis of the PCR reaction, the activated integration was identified based on size and its presence in all YFP-positive embryos with the same expression pattern. The identified integration was sequenced after direct cloning into the TOPO TA cloning kit (Invitrogen, Carlsbad, CA). The genomic integration site was subsequently identified by BLAST against the ENSEMBL zebrafish genome sequence (www.ensembl.org). A sequence was deemed to flank an enhancer detection insertion if it: (1) was present exclusively in eight YFP-positive larvae with the same expression pattern; (2) contained both LTR and linker sequence; (3) matched to a genomic sequence starting within three bases downstream of the LTR; (4) showed 95% or greater identity to genomic DNA; and (5) matched to only one locus with 95% or greater identity (modified after Wu et al., 2003).

In situ hybridization and immunodetection

In situ hybridizations were carried out as described (Jowett and Lettice, 1994). For immunodetection of YFP, Embryos were fixed for 3 hours at room temperature in 4% PFA.

After rinsing three times for 10 minutes in PBT, embryos were dehydrated stepwise: twice for 5 minutes in 50% methanol/50% PBT; twice for 5 minutes in 100% methanol and stored in 100% methanol overnight at –20°C. After stepwise rehydration into PBT, embryos were permeabilized in 0.01 mg/ml Proteinase K (in PBT) for 5 minutes at room temperature then rinsed three times for 5 minutes in PBT.

Post fixation was performed in 4% PFA for 20 minutes at room temperature then embryos were rinsed three times for 5 minutes in PBT. Preparations were blocked at room temperature for 2 hours in incubation buffer (10% goat serum, 1% DMSO, 1% Triton-X in PBS).

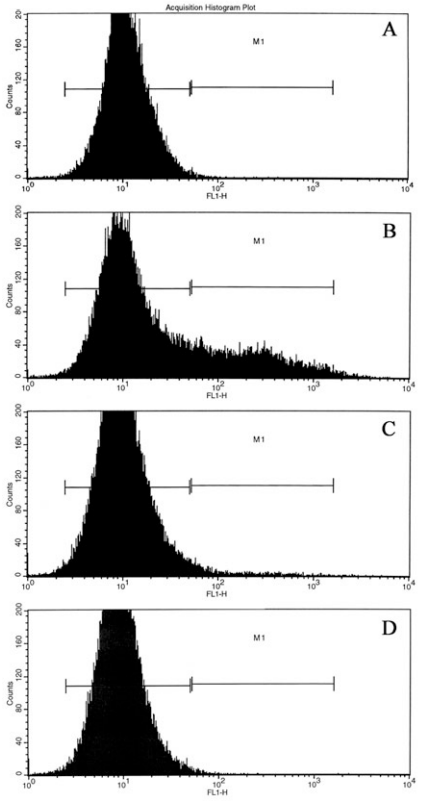
Incubation with primary antibody was carried out for 24 to 30 hours at 4°C (polyclonal rabbit anti GFP from Torrey Pines Biolabs at 1:000 dilution in incubation buffer).

After six 30 minutes rinses in PBT at room temperature, specimens were blocked for 2 hours in incubation buffer. Incubation with secondary antibody was carried out for 18 hours at 4°C (Sigma goat ant-rabbit IgG, catalogue number A11034, at 1:200 dilution in incubation buffer), followed by a six 30-minute rinses at room temperature. The HRP signal was then developed by incubation in DAB solution/0.3% H_2O_2 for 20 until the reaction was stopped by five 5-minutes washes in PBS.

Results

A fraction of integrated proviruses are activated in zebrafish cells

A zebrafish proximal GATA2 promoter fragment of 1024 bp (Meng et al., 1997) was ligated into the MLV-derived CL vector (Naviaux et al., 1996) upstream of a reporter gene



	% YFP positive cells	Average number of provirus integrations per cell.	Fraction of integrated provirus that are transcriptionally activated
A	-	-	-
B	22.93	1.62	0.14
C	2.54	0.20	0.13
D	0.39	0.013	0.3

Fig. 1. FACS analysis of zebrafish fibroblast cells (Pac2) infected with serial dilutions of the viral vector CLGY. Cells were infected with pseudotyped virus. Two days post infection, cells were harvested and analyzed for YFP expression by flow cytometry using a FACSCalibur Flow Cytometry Analyzer (Becton Dickinson, USA). (A) Uninfected cells. (B) Cells infected with concentrated CLGY stock (1.1×10^8 CFU/ml). (C) Cells infected with 1:10 dilution of concentrated virus stock. (D) Cells infected with 1:100 dilution of concentrated virus stock. To calculate average number of provirus integrations per cell, real-time quantitative PCR was performed on genomic DNA isolated from infected cells. Genomic DNA from cells containing two copies of an integrated provirus per cell was used as reference for provirus number, while variations in template amount were normalized against an endogenous target sequence (Chen et al., 2002).

encoding yellow fluorescent protein (YFP), thus creating CL-GATA2-YFP (CLGY).

Upon insertion into human and mouse genomes, the MLV long terminal repeat (LTR) functions as a strong promoter driving viral transcription. By contrast, in cultured zebrafish cells and in embryos, we observed that the enhancer/promoter in the LTR of MLV-derived vectors has very low, if any, activity. We infected cultured zebrafish Pac2 fibroblasts (Chen et al., 2002) with concentrated CLGY virus (Fig. 1). Subsequent fluorescence-activated cell-sorting analysis showed that concentrated virus yielded about 22% YFP-expressing cells (Fig. 1B), whereas real-time quantitative PCR (qPCR) analysis of these cells proved that the average infection rate was 1.62 integrations per cell (Fig. 1B). Hence, 13-14% of integrations (Fig. 1) resulted in expression of YFP in Pac2 cells, whereas the remainder was silent under these conditions. Subsequent ten-fold and 100-fold dilutions of the virus demonstrate that this activity diluted roughly in a linear manner (Fig. 1C,D). By contrast, infection of human HEK293 cells with comparable titers resulted in virtually every cell expressing YFP (data not shown). As the viral LTR is not an active promoter in zebrafish cells, we hypothesized that in those cases where the fish cells expressed YFP, the GATA2 promoter was activated by *cis*-regulatory elements in the cellular genome. Virus injected zebrafish embryos were inspected by fluorescence microscopy at 24 and 48 hours post fertilization (hpf). We found mosaic expression of YFP throughout the body, with varying levels of expression (see Fig. S1 in the supplementary material), whereas embryos injected with the promoterless, but otherwise identical, virus vector CLY showed no expression of YFP (Fig. S1 in the supplementary material). This suggests that in zebrafish embryos, expression of the YFP reporter depends on the presence of the GATA2 basal promoter, and that those cases where expression is detected represent somatic enhancer detection events.

Regulated expression in transgenic lines

To investigate if integrated proviruses could be activated to express YFP after passing through the germline, injected fish

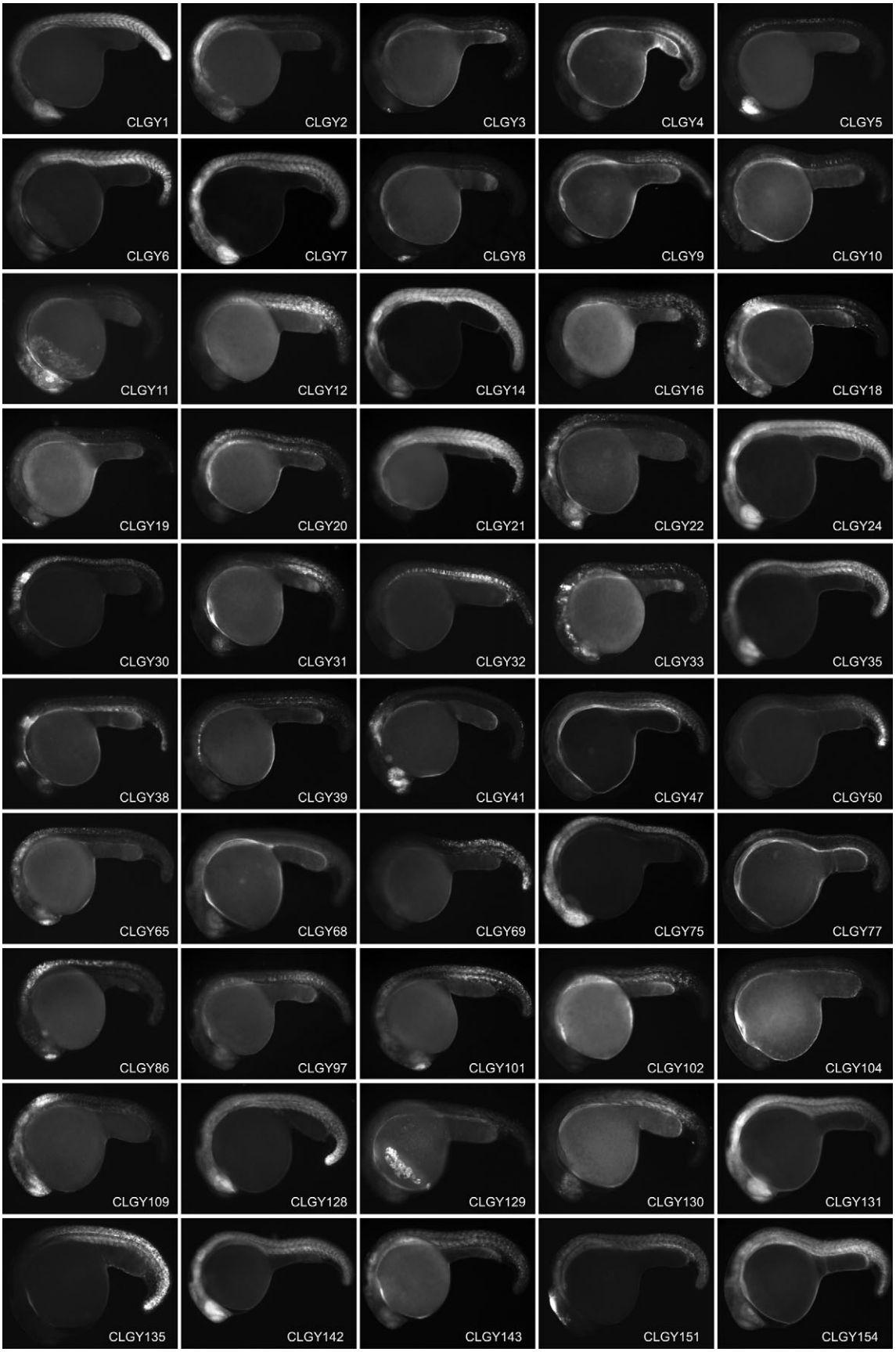


Fig. 2. For legend see p. 3804.

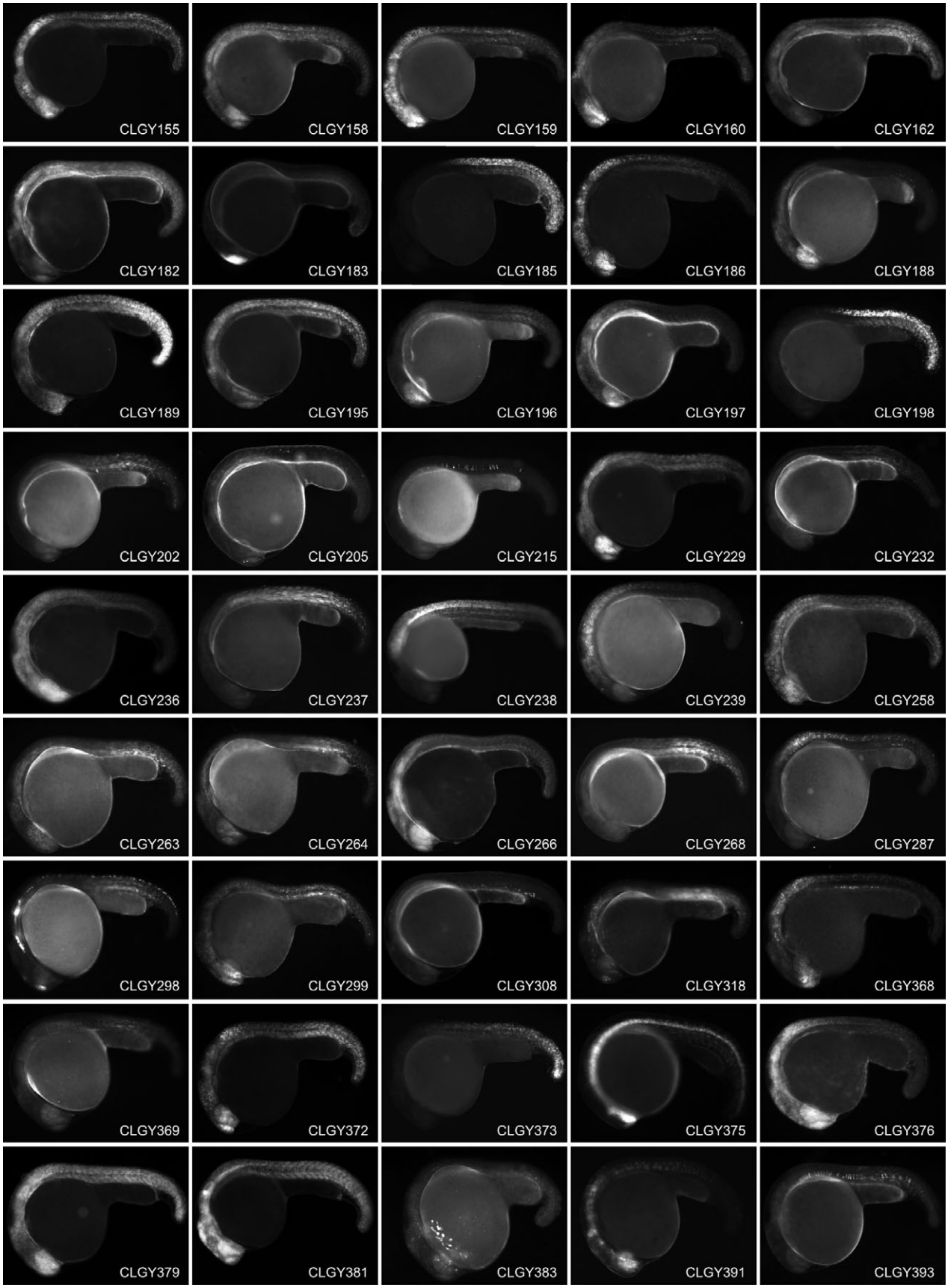


Fig. 2 continued. For legend see p. 3804.

embryos were raised to sexual maturity and crossed to nontransgenic wild-type fish. The F1 progeny from these crosses were then screened with a fluorescence microscope at one-day post fertilization (1 dpf). We found that on average one out of three founders transmitted an activated insertion, leading to the isolation of 95 individual reporter expression patterns. As expected, we observed non-Mendelian inheritance of activated provirus in the F1 clutches, where frequencies of YFP-expressing embryos for any given expression pattern were in the range of 1-20%. These rates reflect the late viral infection of germline cells, and have been reported previously (Gaiano et al., 1996a; Amsterdam, 2003). The number of different activated insertions transmitted through the germline of positive founders was characteristically one, but there were cases of four different patterns from the same founder. In each case, all positive embryos with one distinct expression pattern were collected and grown to sexual maturity to create a transgenic line of fish. From F1 onwards, activated insertions were inherited in a Mendelian fashion, and we have observed stable expression up to generation F5 for multiple transgenic lines (data not shown). We found expression in distinct patterns (Fig. 2; http://clgy.no/clgyimages/), ranging from a subset of cell types (e.g. CLGY19), CNS domains (e.g. CLGY5) to widespread expression (e.g. CLGY14). Enhancer detection events were most frequently observed expressing in the central nervous system (CNS), but also in derivatives of the mesoderm (e.g. somites or notochord: CLGY21 and CLGY32, respectively) or of the endoderm (e.g. hatching glands: CLGY129).

Fig. 2. Stable transgenic lines showing spatially and temporally restricted expression of YFP around 24h post fertilization, anterior is to the left and dorsal to the top. Main domains of expression at 1 dpf (CLGY number is in parentheses): (1) Widespread expression, most pronounced in hypothalamus and posterior somites. (2) Diencephalon, hindbrain and anterior spinal cord. (3) Olfactory placodes. (4) Yolk syncytial layer, ventral caudal mesoderm. (5) Retina, epiphysis and spinal cord. (6) Developing skeletal muscle. (7) Widespread in CNS and muscle. (8) Dorsal telencephalon. (9) Hindbrain and posterior mesenchymal cells. (10) Subset of cells in CNS and notochord. (11) Neural crest and cranial ganglia. (12) Differentiating skeletal muscle and notochord. (14) Widespread, pronounced in telencephalon, retina, hindbrain and most posterior tissues. (16) Tail tip, notochord and skeletal muscle. (18) CNS, hindbrain (19) Dorsal telencephalon and mottled expression in notochord and spinal cord. (20) Hindbrain and spinal chord. (21) Developing skeletal muscle. (22) Telencephalon, olfactory placodes, hindbrain and spinal cord. (24) Widespread expression in CNS and muscle. (30) Hindbrain and spinal cord. (31) Retina and differentiating muscle. (32) Notochord and differentiating muscle. (33) Retina, hindbrain and spinal cord. (35) Widespread expression in CNS and muscle. (38) Retina, mid-hindbrain and hindbrain. (39) Sheath cells of the notochord (41) Retina, telencephalon and hindbrain. (47) Posterior mesenchymal cells. (50) Tail tip. (65) Telencephalon, olfactory placodes, hindbrain and spinal cord. (68) Widespread. (69) Spinal cord, posterior to somite 4/5 boundary. (75) CNS. (77) Spinal cord. (86) Telencephalon, hindbrain and spinal cord. (97) Notochord and differentiating muscle. (101) Telencephalon and differentiating muscle. (102) Differentiating skeletal muscle. (104) Cells in hindbrain and anterior part of spinal cord. (109) Forebrain and spinal cord. (128) Widespread, pronounced in ventral diencephalon and tailbud. (129) Hatching glands. (130) Retina and differentiating muscle. (131) Blood, retina, telencephalon, hindbrain, and muscle. (135) Posterior trunk. (142) Widespread. (143) Skeltal muscle and CNS. (151) Dorsal diencephalon and dorsal mesencephalon. (154) Widespread, strong in retina and telencephalon. (155) Telencephalon, retina, midbrain, rhombomeres, spinal cord. (158) Widespread. (159) Central nervous system and developing muscle. (160) Telencephalon and diencephalon. (182) Widespread in CNS and muscle. (183) Telencephalon. (185) Spinal cord and posterior trunk. (186) Central nervous system. (188) Retina and telencephalon. (189) Widespread. (195) Telencephalon, ventral diencephalon, MHB, hindbrain and spinal cord. (196) Ventral forebrain and retina. (197) Telencephalon and retina. (198) Spinal cord and posterior trunk. (202) Differentiating muscle. (205) Cells in forebrain, hindbrain and spinal cord. (215) Cells in notochord. (229) Widespread, strong in retina and hindbrain. (232) Mesenchymal cells (236) Retina and CNS. (237) Skeletal muscle. (238) Posterior hindbrain and spinal cord. (239) Hindbrain and spinal cord. (258) Widespread expression. (263) Retina and posterior mesenchyme. (264) Weak widespread. (266) Widespread, pronounced in CNS. (268) Differentiating muscle. (287) Hindbrain and spinal cord. (299) Retina, telencephalon and posterior mesenchymal cells. (308) Isolated cells in mesenchyme. (318) Telencephalic cells, neural crest and posterior mesenchyme. (368) Telencephalon, olfactory placodes, hindbrain and spinal cord. (369) Retina, notochord and differentiating muscle. (372) Blood, retina, telencephalon, hindbrain, and muscle. (373) Posterior trunk. (375) Ventral medial CNS. (376) CNS, strongest in retina, hindbrain and anterior spinal cord. (379) Widespread. (381) Widespread, strongest in CNS. (383) Retina and hatching glands. (391) Diencephalon and hindbrain. (393) Notochord. High resolution images are at www.clgy.no/clgyimages.

Genomic mapping of activated insertions

To identify the genomic location of an enhancer detecting proviral insertion, it was necessary to distinguish the activated insertion from the majority of non-activated insertions (see Materials and methods). In all cases shown here, there was only one amplified flanking sequence that segregated with all YFP-positive embryos (see Fig. S2 in the supplementary material), presumably owing to the low total number of insertions (~3/founder) relative to the 25 zebrafish chromosomes. The flanking fragment of each activated provirus was sequenced (see Table S1 in the supplementary material) and used to search the zebrafish genome assembly in the Ensembl database (http://www.ensembl.org/) using BLASTN. For unambiguous mapping of the insertion onto the zebrafish genome, we required a single unique hit with over 95% identity to genomic sequence (Table 1). The average amplicon size was 127 bp, and the shortest, 28 and 25 bp, produced single unique hits of 100% over the entire length (CLGY47 and CLGY198). In 10 out of 95 cases, we did not obtain any high identity matches in the genome database, and in a further 13, there were multiple hits with high identity. In an additional six cases, the insertion was mapped to a contig not yet assembled or found close to the end of an assembled contig, leaving 65 insertions that gave single unique hits in assembled sequence or large unassigned contigs. Over half (33/64) of these insertions are located in finished sequence according to the Sanger Centre genome annotation, usually either BAC sequence or manually annotated sequence. Further improvements in the zebrafish genome sequence will allow resolution of the remaining 31 cases. Alternatively, longer flanking sequence can be generated starting from the sequence already obtained. We have listed the flanking sequences identified in each transgenic line in Table S1 (see supplementary material).

Majority of activated insertions map close to/inside genes

Based on the current annotation of the zebrafish genome, in 27 out of the 33 cases that were mapped to finished sequence, the activated vectors were found to have integrated into or within 15 kb up- or downstream of a transcript. In the remaining six cases, we observed that the closest gene was further away, up to 132.8 kb from the integrated vector (Table 1). Although in the latter case the integration is in a gene desert also found in the human genome, we cannot exclude that, owing to the still unfinished status of the zebrafish genome, genes may be annotated closer to some of these insertion sites in the future. Many of the regions to which insertions were mapped contain genes that encode transcriptional regulators involved in development (Table 1), suggesting that the enhancers of these genes interact with our inserted vector. In other cases we found genes that are not thought to be associated with development but appear to have strong enhancers also interacting with the GATA2 promoter in the detection vector, for example *cyclinD1* (CLGY131 and CLGY372) or *apoeb* (CLGY4 and CLGY162). Both these cases, as well as some of the transcriptional regulators, have published gene expression patterns and we were interested to see whether they match the observed YFP expression patterns.

Reporter gene expression patterns resemble those of endogenous genes close to the insertion site

We obtained antisense in situ probes for 10 candidate genes identified by our approach that are described in the literature and found that the expression of YFP closely resembles the endogenous transcriptional pattern (Figs 2 and 3). In Fig. 3 an immunostain of the YFP pattern was compared with the RNA pattern of the candidate gene mapping close by in the zebrafish genome. Expression patterns were found to be very similar in CLGY5/*pax6.2* (*pax6b* – Zebrafish Information Network) (Nornes et al., 1998), CLGY4/*apoeb* (B. Thisse, S. Pflumio, M. Fürthauer, B. Loppin, V. Heyer, A. Degrave, R. Woehl, A. Lux, T. Steffan, X. Q. Charbonnier, and C. Thisse, unpublished), CLGY183/*emx3* (Houart et al., 2002), CLGY198/*hoxc8a* (Prince et al., 1998), CLGY11/*otx1l* (Hauptmann et al., 2002), CLGY375/*ptc1* (Concordet et al., 1996), CLGY21/*snai1a* (Thisse et al., 1993) and CLGY75/*sox19* (B. Thisse, S. Pflumio, M. Fürthauer, B. Loppin, V. Heyer, A. Degrave, R. Woehl, A. Lux, T. Steffan, X. Q. Charbonnier, and C. Thisse, unpublished). CLGY183 and CLGY375 had been mapped to unfinished scaffold sequence but show the correct pattern. CLGY75 was mapped to finished sequence, but the insertion is found 14 kb downstream of the *sox19* transcriptional unit. Nevertheless, the endogenous RNA pattern is virtually the same as the reporter expression of the transgenic line. Likewise, CLGY11 is an insertion over 30 kb upstream of the *otx1l* transcriptional unit, yet the patterns shown here are closely matching (Fig. 3).

A notable difference between in situ patterns and YFP expression is due to the higher stability of the reporter protein compared with the endogenous mRNA, resulting in protein

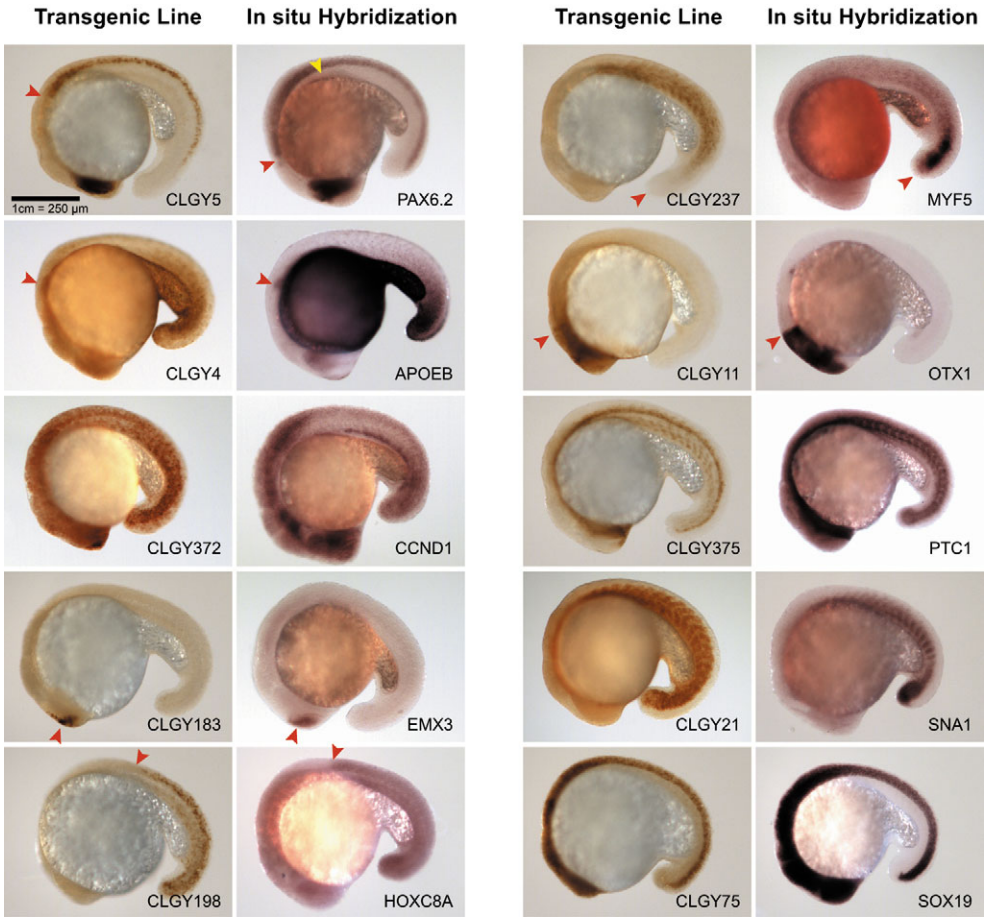


Fig. 3. Comparison of YFP expression in transgenic lines with RNA expression patterns of candidate genes. Embryos are 18 hours post fertilization, anterior towards the left and dorsal upwards. CLGY5/*pax6.2* are expressed in the same domains, retina, pineal and spinal cord, with exception of the pancreas seen in the in situ stain (yellow arrowhead) and much weaker expression in hindbrain in the transgenic line (anterior limit of expression domain is marked with a red arrowhead). CLGY4/*apoeb* are highly similar in expression, including the pectoral fin buds (arrowhead). CLGY372/*ccnd1* are very similar with a widespread and highly dynamic expression. CLGY183/*emx3* are expressed in the telencephalon only (arrowhead). CLGY198/*hoxc8a* share the same anterior boundary (arrowhead). Expression in CLGY237/*myf5* is similar in the developing muscle, but the tail bud expression (arrowhead) of the endogenous RNA is not seen in the transgenic line. CLGY11/*otx1l* have a highly similar expression pattern, anterior boundary shown by an arrowhead. CLGY375/*ptc1* are very similar. CLGY21/*snai1a* are expressed in the same domain, skeletal muscle. CLGY75/*sox19* are both expressed at high levels throughout the CNS.

Table 1. Characterization of genomic location of transcriptionally activated provirus

CLGY number	Genomic position relatively to nearest transcript	Chromosome	Nearest gene	Ortholog prediction of nearest gene to human (if not noted otherwise)
1	Donor splice site, intron 1.	9	pl10	Dead-box protein 3 (DDX3X)
2	nd4	nd	–	–
3	395 bp downstream	23	Novel prediction	her4
4§	5649 bp downstream	16	Apoeb	Apolipoprotein A-I precursor (APOA1)
5§	2nd intron	7	Pax6b	Paired box protein PAX-6 (PAX6)
6¶	4727 bp upstream	5	Tbx16	T-box transcription factor TBX6 (TBX6)
7	nd4	nd	–	–
8¶	In exon	16	Genscan prediction	nd
9	2692 bp upstream	20	Novel prediction	NHS-like 1 (NHSL1)
10	nd2	nd	–	–
11§	32,296 bp upstream.	1	Otx1I	nd
12	814 bp downstream	Zv4 NA6752	Novel prediction	Chromosome 10 open reading frame 104 (C10orf104)
14	nd4	nd	–	–
16	1st intron	15	Hsp47	Serpin I2 precursor (SERPINI2)
18	1st intron	16	Genscan prediction	Odd Oz/ten-m homolog 3 (ODZ3)
19	7470 bp upstream	22	Novel prediction	ATP synth. delta chain, mitochondrial prec. (ATP5D)
20§	1360 bp upstream	18	Rara2a	Retinoic acid receptor beta (RARB)
21§	14,041 bp downstream	11	Snai1a	Zinc finger protein SLUG (SNAI2)
22§	70,441 bp upstream	20	Sox11b	Transcription factor SOX-11 (SOX11)
24§	3153 bp upstream	16	Pou5f1	POU domain, class 5, transcription factor 1 (POU5F1)
30	nd2	Zv4 NA16664	–	–
31	1st intron	20	Zgc:56602	Hypothetical protein FLJ11712
32§	6th intron	22	Genscan prediction	nd
33	nd4	9	Genscan prediction 00000003859	Smad-interacting protein1b
35*,¶	8084 bp upstream	16	Pou5f1	POU domain, class 5, transcription factor 1 (POU5F1)
38	10th intron	22	Novel prediction	RAB GTPase activating protein 1 (RABGAP1)
39	36,022 bp downstream	19	Novel prediction	Similar to CG6405 gene product (NM_145269).
41§	5890bp upstream	22	Novel prediction	Transducin-like enhancer protein 1 (TLE1)
47§	1st intron	3	Novel prediction	Glucose-6-phosphatase (G6PC)
50	1st intron	Zv4 NA12687	Genscan prediction	nd
65§	132,831 bp upstream	20	Sox11b	Transcription factor SOX-11 (SOX11)
68	4th intron	Zv4 NA17998	Genscan prediction	nd
69¶	3′UTR	Zv4 scaffold335	Hoxa9a	Homeobox protein HOX-A9 (HOXA9)
75†,§	14 027 bp downstream	5	Sox19	Transcription factor SOX-19 (SOX19)
77	884 bp upstream	18	Gro2	Transducin-like enhancer protein 3 (TLE3)
86	69,259 bp upstream	14	Novel prediction	Potential phospholipid transporting ATPase IG (ATP11C)
97	5th intron	6	Novel prediction	Probable urocanate hydratase (HUTU_HUMAN)
101	10,487 bp downstream	13	Q804r3	ORF2 of novel retrotransposon.
102§	2nd intron	2	Novel prediction	nd
104	nd4	nd	–	–
109	nd1	nd	–	–
128‡,§	257 bp down-/2634bp upstream	8	Novel/foxd4	COBW-like protein (NM_018491)/Forkhead box D4
129	nd[4]	nd	–	–
130	2nd intron	1	Efnb2b	Ephrin-B2 precursor (EFNB2)
131§	2675 bp upstream	24	Ccnd1	G1/S-specific cyclin D1 (CCND1)
135	nd4	nd	–	–
142	nd4	nd	–	–
143	nd4	nd	–	–
151	2335 bp upstream	10	Tpbgl	Trophoblast glycoprotein (TPBG)
154	nd2	13	–	–
155	nd1	nd	–	–
158	nd1	nd	–	–
159	nd4	nd	–	–
160	nd1	nd	–	–

detection in cells whose precursors expressed the message earlier, especially in cases where mRNA expression is highly dynamic, e.g. CLGY372/*ccnd1*. In addition, protein expression can be expected to lag behind transcription, leading to differences between in situ and protein expression. For example, *snai1a* mRNA is detected mostly in the tip of the tail, whereas the reporter is found to accumulate in skeletal muscle, the progeny of the cells that expressed the message earlier. In some cases, aspects of the endogenous pattern were missing from the enhancer detection lines. For example, *pax6.2* is expressed strongly in the pancreas (Fig. 3, arrowhead) (B. Thisse, S. Pflumio, M. Fürthauer, B. Loppin, V. Heyer, A. Degrave, R. Woehl, A. Lux, T. Steffan, X. Q. Charbonnier, and C. Thisse, unpublished), but this is not found in the insertion line CLGY5 (although this insertion is located inside the gene). In addition, expression in the hindbrain is much weaker in CLGY5. These differences are probably not due to lack of detection, but rather reflect lack of interaction of the pancreas and hindbrain enhancer elements with the GATA2 promoter in the proviral insertion. Similarly, *myf5* is expressed in the tail bud (Fig. 3), but YFP is not detected in the tailbud in CLGY237, while the expression in muscle is detected in both transgenic line and in situ hybridization. Thus, although we have not investigated expression patterns for all the

Table 1. Continued

CLGY number	Genomic position relatively to nearest transcript	Chromosome	Nearest gene	Ortholog prediction of nearest gene to human (if not noted otherwise)
162§	3001 bp downstream	16	Apoeb	Apolipoprotein A-I precursor (APOA1)
182	nd4	nd	–	–
183	2899 bp downstream	14	Emx3	nd
185§	4902 bp upstream	23	Hoxc8a	Homeobox protein HOX-C8 (HOXC8)
186	2208/4635 bp downstream[4]	14	Genscan prediction	Zinc finger protein 36, C3H type-like 1 (ZFP36L1)
188	nd1	nd	–	–
189	57,383 bp downstream	9	Zgc:55421	Protein tyrosine phosphatase like (PTPLB)
195	nd2	Zv4 NA3586	–	–
196§	5th intron	4	Genscan prediction	SRY-family
197	nd2	Zv4 NA13949	–	–
198§	3988 bp upstream	23	Hoxc8a	Homeobox protein HOX-C8 (HOXC8)
202§	2nd intron	24	Novel prediction	Mitogen-activated protein kinase 3 (MAP3K3)
205§	222,949 bp upstream	20	Sox11b	sox11
215	1st intron	5	Genscan prediction	nd
229	22,927 bp upstream	nd	Fgfr1	Fibroblast growth factor receptor 1 (FGFR1)
232	1st intron	4	Zgc:55542	Protein phosphatase 1, regulatory subunit 3B (PPP1R3B)
236	5th intron	22	Novel prediction	nd
237§	3970 bp upstream	4	Myf5	Myogenic factor MYF-5 (MYF5)
238§	73 bp downstream	12	Hoxb1b	Homeobox protein HOX-B1 (HOXB1)
239§	4th intron	5	Zgc:55984	Branched-chain a-ketoacid dehydrogenase kin (BCKD)
258	nd1	nd	–	–
263	175 bp upstream	20	Novel prediction	Mediator of DNA damage checkpoint 1 (MDC1)
264§	5th intron	20	Novel prediction	P-selectin precursor (SELP)
266	nd4	nd	–	–
268	nd1	nd	–	–
287§	4th intron	11	Novel prediction	nd
298§	4906 bp downstream	4	Ube2h	Ubiquitin conjugating enzyme E2H (UBE2H)
299	nd1	nd	–	–
308	nd1	nd	–	–
318§	1st intron	10	Cldn7	Claudin-7 (CLDN7)
368§	89,331 bp upstream	20	Sox11b	Transcription factor SOX-11 (SOX11)
369§	2nd intron	10	Zgc:77101	Ubiquitin-like 3 (Ubl3)
372§	743 bp upstream	24	Ccnd1	G1/S-specific cyclin D1 (CCND1)
373	nd1	nd	–	–
375¶	348 bp upstream	2	Ptc1	Patched protein homolog 1 (PTCH)
376§	74,656 bp downstream	21	Novel prediction	Double-stranded RNA-binding zinc finger protein 346 (NM_012279)
379	3rd intron	9	Genscan prediction	Adenylate cyclase 6, isoform a (ADCY6)
381	nd2	Zv4 NA14257	–	–
383§	2651 bp downstream	20	Genscan prediction	nd
391	nd4	nd	–	–
393	5th intron	20	Novel prediction	Gamma-tubulin complex component 2 (TUBGCP2)

Flanking sequences of activated proviruses were obtained using LIM-PCR (Wu et al., 2003). Genomic position of the integrated vector was identified by BLASTN of the flanking sequence against the Ensembl zebrafish genome browser (zebrafish whole genome shotgun assembly sequence version 4, Zv4) (http://www.ensembl.org/).

*CLGY35 is closer to a novel gene (6353 bp upstream), but shows the expression pattern of *pou5f1*.

†CLGY75 is closer to a novel gene (11,148 bp upstream), but shows the expression pattern of *sox19*.

‡CLGY128 is in between *foxd4* and a novel gene, closer to the latter. The expression pattern is consistent with a forkhead domain factor.

§Integrations mapped to finished sequence.

¶Cases in unfinished scaffold sequence that could be confirmed by in situ pattern.

Abbreviations: nd, not determined; nd1, no hits in Zv4 assembly (10); nd2, provirus integrated in non-assembled sequence (6); nd4, multiple hits with at least 95% identity (13).

Genscan predictions are supported by EST data in ENSEMBL.

candidate genes near the insertions presented in this work, we conclude that in many cases the expression pattern of a gene close to the insertion site can be used to confirm the genome data. For example, CLGY183 and CLGY375 were mapped to scaffold (unfinished) sequence but the expression patterns could be verified, thus confirming the mapping data. A further two insertions, although not located in finished sequence, could be confirmed through in situ hybridization patters from the literature: CLGY24/*pou5f1* (Burgess et al., 2002) and CLGY6/*tbx16* (B. Thisse, S. Pflumio, M. Fürthauer, B. Loppin, V. Heyer, A. Degrave, R. Woehl, A. Lux, T. Steffan, X. Q. Charbonnier, and C. Thisse, unpublished). CLGY8 is an insertion into a novel gene represented by an EST according to ENSEMBL; we also confirmed this expression pattern using in situ hybridization (data not shown).

By contrast, CLGY298 (Fig. 4), which is located in finished sequence, does not resemble the expression pattern of either of the two genes flanking it (not shown). The gene it is closest to, *ube2h*, is broadly expressed throughout the embryo (H.K. and T.S.B., unpublished), whereas *nrf1*, about 15 kb away on the other side, is expressed throughout the central nervous system (Becker et al., 1998). Reporter expression in CLGY298, however, is found in sensory neurons in the olfactory placodes, the trigeminal ganglia, the inner ear, the lateral line, the spinal

cord and the retina, and continues to be expressed there at least up to day 10 (Fig. 4; not shown). There are also three genes encoding miRNAs between *nrf1* and *ube2h*, which are conserved from human to fish. These miRNAs are 11, 10 and 9.8 kb from CLGY298. The corresponding genomic region contains many transcriptional units, and perhaps the basal promoter in the integrated provirus is driven by enhancers other than the ones of the two flanking genes, for instance *plexin A4* (*plxna4* – Zebrafish Information Network) maps far downstream of *nrf1* and is expressed in primary sensory neurons (Miyashita et al., 2004). We conclude that in the majority of our enhancer detection lines the reporter expression is closely matching the endogenous pattern of a nearby gene, but there can be exceptions where the pattern can be that of a gene that is not the closest to the insertion. However, in three cases described here, CLGY35/*pou5f1*, CLGY75/*sox19* and CLGY128/*foxd5*, the expression pattern of the other candidate gene is not known and it is possible that both genes are regulated by the same enhancer(s), forming what has been termed a regulatory landscape (Spitz et al., 2003).

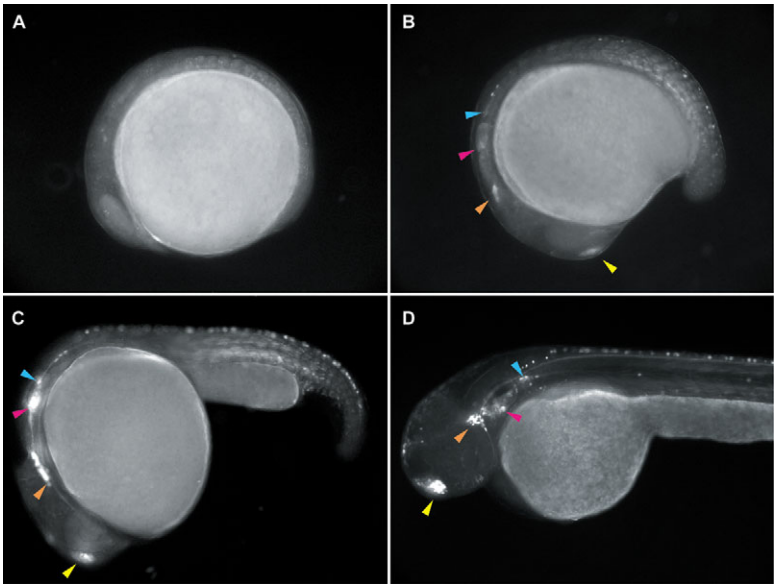


Fig. 4. Expression in sensory placodes in CLGY298. (A) 8 hpf; (B) 16 hpf; (C) 24 hpf; (D) 48 hpf. Anterior is towards the left. Expression of YFP is first seen at 16 hpf in the olfactory placode (yellow arrowhead), the trigeminal placode (orange arrowhead), the inner ear (pink arrowhead) and the lateral line primordium (blue arrowhead). At 24 hours the latter three primordia migrate together before reaching their final destination, as seen at 48 hours (D). Expression is also seen in skeletal muscle early on, as well as in sensory neurons in the spinal cord.

Allelic integrations

Wu et al. have reported that of the 903 MLV randomly sequenced integrations in the human genome, no integration hot spots had been observed, concluding that the resolution of this number of integrations may not allow the observation of obvious integration preferences of MLV vectors. We have screened an estimated 800 insertions in the zebrafish genome, and of 65 activated insertions mapped here, four loci were hit twice: CLGY4/162, CLGY24/35, CLGY131/372 and CLGY185/198 (Table 1). CLGY4 and CLGY162 are 5649 bp and 3001 bp downstream of the last exon of *apoeb*, CLGY24/35 are insertions 3153 bp and 8084 bp upstream of *pou5f1*, and CLGY131/372 are insertions 2675 bp and 743 bp upstream of the start codon of *ccnd1*, a G1/S-specific cyclin. The CLGY185/198 alleles represent integrations 4902 bp and 3988 bp upstream of the *hoxc8a* gene (Table 1) Each of these allelic pairs of integrations have very similar if not identical expression patterns (Fig. 2), suggesting that in each case, the integrations have landed in an area controlled by the same enhancer(s). It is striking that in each of the five cases the integrations are close to a gene in a somewhat restricted area. It is not clear whether these genomic locations represent areas in the zebrafish genome that are more accessible to viral integration, or whether the selection for activated insertions biases towards identifying integrations in regions with a higher probability of activation. We conclude that it is possible that MLV vectors do have preferred integration regions with respect to where in or around a gene the insertion occurs. Certainly, our screen creates a bias towards integrations that will be expressed, and therefore around genes that have certain types of enhancers (or around the enhancers themselves), but higher numbers of insertions are required to resolve this issue.

Integrations in a gene desert

We identified four integrations in a genomic region on chromosome 20 spanning a 320 kb interval without known or predicted genes, flanked on one side by a novel transcript, a putative orthologue to human allantoicase (ALLC, ENSDART00000033971) and on the other side by the zebrafish *sox11b* gene. We observed that these insertions exhibit very similar expression patterns in olfactory placodes, dorsal telencephalon and hindbrain (Table 1 and Fig. 2; CLGY22, CLGY65, CLGY205 and CLGY368). The sites of expression of *sox11b* at 24 hpf include the telencephalon and hindbrain (Rimini et al., 1999; De Martino et al., 2000), while the expression pattern for the novel zebrafish transcript, to which CLGY205 is closer than to *sox11b*, is not known. CLGY205 is the insertion farthest from *sox11b* (roughly 222kb), and there are fewer cells expressing YFP in the telencephalon and hindbrain in CLGY205 compared with the other three insertions. A gene desert is also found on human chromosome 2 upstream of *SOX11*, and an ultra conserved region (UCR) was mapped to this region (Sandelin et al., 2004; Woolfe et al., 2005). This UCR maps 89 kb upstream of *sox11b*, and CLGY205 is therefore farthest away from it. Whether this UCR contains the enhancer responsible for reporter gene expression in these four insertions, however, remains to be seen. Regardless of this, these cases demonstrate that our methodology will be useful to identify vertebrate gene deserts that contain strong enhancers.

Discussion

We show here that a retroviral vector based on an MLV genome acts as an efficient enhancer detector in the zebrafish genome. Although enhancer detection has been reported in vertebrates

previously (Korn et al., 1992; Bayer and Campos-Ortega, 1992; Grabher et al., 2003; Balciunas et al., 2004; Parinov et al., 2004), our retroviral vector allows efficient integration rates and recovery of transgenic lines on genome scale: The average rate of identifying an enhancer detection event in F1 progeny infected with CLGY virus was one in three founders. Therefore, the generation of activated insertions is not rate limiting. In one week, one researcher can inject 3000 founders, corresponding to 8000 genomic insertions and to 1000 enhancer detection events. Furthermore, it is possible to increase the viral titer by establishing a stable producer clone for the viral vector (Chen et al., 2002). We currently estimate an average of three insertions per founder while viral titers producing 30 insertions per fish and germline have been reported (Chen et al., 2002). In our current large-scale screen, we find that somewhat lower numbers of integrations (ideally eight per injected founder, corresponding to one enhancer detection event) are easier to handle, as multiple enhancer detection events are more difficult to sort out and because at low insertion number the non-activated integrations segregate out in one to two generations, allowing easy cloning of flanking sequence from single embryos. However, in the future, numbers of integrations 10-fold higher (10,000 activated events/person/week) could allow saturating the zebrafish genome for specific types of patterns or genes in a mid size laboratory. Our enhancer detection system also allows large-scale cloning of flanking sequence and mapping of the integration sites, as retroviral vectors insert as single copies.

The integration preferences of MLV-derived viruses in the human genome were described by Wu et al. (Wu et al., 2003), showing a certain tendency of MLV to integrate close to transcription start sites of genes. About 34% of mapped MLV integrations had occurred in RefSeq genes, while a further 11.2% integrated within 5 kb upstream of genes (Wu et al., 2003). However, in our study, only about one in eight integrations were activated, suggesting that either these vectors integrate into genes less frequently in the zebrafish genome, or, most likely, that not all integrations into genes are expressed during embryogenesis. Similar to MLV vectors, P-elements preferentially integrate into 5′ regions of genes in *Drosophila* (Bellen, 1999), but P-element mediated enhancer detection has a frequency of activation five times higher than in our study (Bellen, 1999; Bier et al., 1989; Wilson et al., 1989; O'Kane and Gehring, 1987). Although the zebrafish genome harbors more genes, it is less compact than the *Drosophila* genome, and therefore the lower enhancer detection frequency might be a function of greater distances between genes, and therefore between cis-regulatory elements. As we have shown here, many activated insertions are within a 15 kb distance, typically less, from the nearest gene, but whether this correlates with the 'striking distance' of the average enhancer is impossible to say as insertions might occur with higher frequency around enhancer sequences. A glimpse of how such distances could be estimated in the future is perhaps given by the four integrations in a gene desert on chromosome 20. Although the three insertions closer to the *sox11b* transcriptional unit have very similar expression patterns, CLGY202, which is 215 kb upstream of the gene, exhibits much weaker expression. A gene desert is also located upstream of the human *SOX11* gene, and was recently found to contain ultra conserved elements (Sandelin et al., 2004; Woolfe et al., 2005). Whether it is these elements that drive expression from the four integrations listed here will have to be confirmed by experiment but the fact that interaction can occur over such large distances suggests that genes with such elements are large targets and may be overrepresented in future enhancer detection screens. This may also be true for insertions into Hox clusters, which are overrepresented in our screen. It is of concern for future large-scale screens that four insertions out of 95 would have landed in the same gene desert; however, an intense search for this pattern in our current screen has revealed only one additional insertion in this gene desert in over 900 additional lines screened (H.K. and T.S.B., unpublished).

A further eight out of 65 insertions listed in this paper are near genes associated with UCRs, and these are the four hox insertions, CLGY5 (*pax6.2*), CLGY11 (*otx1l*), CLGY77 (*TLE3*) and CLGY375 (*ptc1*) (Sandelin et al., 2004; Woolfe et al., 2005). Among the other known genes in this paper, there are many regulators of early development. Interestingly, the basal promoter used in this study is of the *gata2* gene, which is itself an early developmental regulator and is associated with UCRs (Sandelin et al., 2004). Whether the choice of promoter has any bearing on the types of genes identified in enhancer detection screens, however, remains to be seen.

In the absence of knowing the expression patterns of all candidate genes in the vicinity of an insertion, we have used the gene closest to the insertion to predict the specificity of the cis-regulatory element(s) driving expression of the YFP reporter. This is an approximation, owing to current limited knowledge of the location of cis-regulatory sequences, and we have shown that this is not always correct. For example CLGY75 is 14 kb upstream of *sox19* but is closer to another gene (ENSDARG00000034116), yet displays the expression pattern of *sox19* (Fig. 3). It is, however, possible that both genes are regulated by the same sequences, similar to the *hoxd* regulatory landscape described by Spitz et al. (Spitz et al., 2003). There is also a detection event not resembling the expression patterns of genes located on either side of it: CLGY298 shows a very specific expression pattern in sensory placodes, but the closest gene, *ube2h*, has a much more widespread expression pattern (H.K. and T.S.B., unpublished), and the next gene on the other side, *nrf1*, also shows a different expression pattern (Becker et al., 1998). To further complicate matters, there are also three miRNAs between these two genes, and *plexin A4*, far beyond *nrf1*, has an expression pattern reminiscent of CLGY298 (Miyashita et al., 2004).

In this case, a possible explanation might be that the basal promoter of our vector is not compatible with the enhancers of the neighboring genes, and/or is activated by different regulatory elements, for example those of *plexin A4*.

Based on genomic mapping, two of the integrations reported here could disrupt gene expression: CLGY8 and CLGY69 have occurred into an exon and 3′UTR, respectively. Furthermore, seven of the transgenic lines carry insertions within the first 2 kb upstream of the transcript and another five are located in the 1st intron (Table 1, Fig. 4), locations with a tendency to be mutagenic as shown in insertional genetic screens in zebrafish (Amsterdam, 2003; Amsterdam et al., 1999). So far one of these (CLGY375, *ptc1*) has turned out to be homozygous lethal (S.E. and T.S.B., unpublished). By cloning the flanking sequences and by mapping of activated insertions to the genome, one can predict the potential mutagenicity of the

insertion before screening for phenotypes, including subtle and non-lethal adult phenotypes that would be missed in a conventional phenotype-driven screen. However, based on our numbers, we estimate that only 5-10% of enhancer detection insertions will result in disruption of gene function, and these will not necessarily have a detectable phenotype.

We have shown here that many insertions can be mapped to the zebrafish genome sequence, rapidly generating candidate genes whose expression patterns can be compared with that of the enhancer detection line. Our forward screen makes it feasible to assign expression profiles and function to large numbers of genomic loci and associated novel and predicted transcripts, as well as putative cis-regulatory sequences, in particular the recently identified UCRs. About 1.7% of the human genome are conserved non-genic sequences (CNGs; Dermitzakis et al., 2005). A subset of these probably have cis-regulatory function during embryonic development and can be identified by sequence conservation and tested through transgenic approaches (e.g. Nobrega et al., 2003; Woolfe et al., 2005). However, not all regulatory sequences are highly conserved, and our approach represents the first systematic attempt to characterize genomic regions for their cis-regulatory activity in any vertebrate species.

Collections of transgenic lines with similar or overlapping expression patterns can serve as a starting point for characterization of developmental signaling pathways in particular tissues or organs for isolation and testing of cis-regulatory sequences that confer similar or identical expression patterns. In *Drosophila*, a major breakthrough of P-element mediated enhancer detection was the establishment of specific markers for tissues that had been previously difficult to visualize (Bellen, 1999; Bier et al., 1989; Wilson et al., 1989; O'Kane and Gehring, 1987). Retroviral vector mediated enhancer detection in the zebrafish opens up the possibility of studying cellular origin and lineages in any tissues and organs. For example, the vertebrate brain is made up of a large number of different neurons of which many are not well described. With the large number of transgenic lines that can be generated using this approach it should be possible in the future to label a large part of the different categories of neurons present in the vertebrate brain.

This work was supported a Sars Centre core grant, by the National Programme in Functional Genomics in Norway (FUGE) in the Research Council of Norway and by additional funding from the University of Bergen to T.S.B. Our current large-scale screen is also supported by the European Commission as part of the ZF-Models Integrated Project in the 6th Framework Programme (Contract No. LSHG-CT-2003-503496). We thank Drs S. M. Burgess, I. Verma, L. Bally-Cuif, M. Brand, V. Prince, P. Ingham, G. Hauptmann and B. and C. Thisse for reagents, and Drs Shawn M. Burgess, Daniel Chourrout and Julien Ghislain for critical reading of an earlier version of the manuscript. We appreciate critical comments on the manuscript by Dr Adam Amsterdam. Additional thanks go to Jennifer Reagan, and Drs Shawn M. Burgess, Melina E. Hale and Mario Caccamo for technical advice. We gratefully acknowledge expert technical help in our zebrafish facilities by Eilen Myrvold, Sara Ferreira, Merete Nilsen, Tore Samuelsen and Heikki Savolainen. This work is dedicated to the memory of José A. Campos-Ortega.

Supplementary material

Supplementary material for this article is available at
http://dev.biologists.org/cgi/content/full/132/17/3799/DC1

References

Amsterdam, A. (2003). Insertional mutagenesis in zebrafish. *Dev. Dyn.* **228**, 523-534.

Amsterdam, A., Burgess, S., Golling, G., Chen, W., Sun, Z., Townsend, K., Farrington, S., Haldi, M. and Hopkins, N. (1999). A large-scale insertional mutagenesis screen in zebrafish. *Genes Dev.* **13**, 2713-2724.

Amsterdam, A., Nissen, R. M., Sun, Z., Swindell, E. C., Farrington, S. and Hopkins, N. (2004.). Identification of 315 genes essential for early zebrafish development. *Proc. Natl. Acad. Sci. USA* **101**, 12792-12797.

Austin, C. P. and Cepko, C. L. (1994). Retrovirus mediated gene transduction into the vertebrate CNS. *Gene Ther.* **1**, S6-S9.

Balciunas, D., Davidson, A. E., Sivasubbu, S., Hermanson, S. B., Welle, Z. and Ekker, S. C. (2004). Enhancer trapping in zebrafish using the Sleeping Beauty transposon. *BMC Genomics* **5**, 62.

Bayer, T. A. and Campos-Ortega, J. A. (1992). A transgene containing lacZ is expressed in primary sensory neurons in zebrafish. *Development* **115**, 421-426.

Becker, T. S., Burgess, S. M., Amsterdam, A. H., Allende, M. L. and Hopkins, N. (1998). Not really finished is crucial for development of the zebrafish outer retina and encodes a transcription factor highly homologous to Nuclear Respiratory Factor-1 and avian Initiation Binding Repressor. *Development* **125**, 4369-4378.

Bellen, H. J. (1999). Ten years of enhancer detection: lessons from the fly. *Plant Cell* **11**, 2271-2281.

Bier, E., Vaessin, H., Shepherd, S., Lee, K., McCall, K., Barbel, S., Ackerman, L., Carretto, R., Uemura, T., Grell, E. et al. (1989). Searching for pattern and mutation in the Drosophila genome with a P-lacZ vector. *Genes Dev.* **3**, 1273-1287.

Burgess, S., Reim, G., Chen, W., Hopkins, N. and Brand, M. (2002). The zebrafish spiel-ohne-grenzen spg. gene encodes the POU domain protein Pou2 related to mammalian Oct4 and is essential for formation of the midbrain and hindbrain, and for pre-gastrula morphogenesis. *Development* **129**, 905-916.

Burns, J. C., Friedmann, T., Driever, W., Burrascano, M. and Yee, J. K. (1993). Vesicular stomatitis virus G glycoprotein pseudotyped retroviral vectors: concentration to very high titer and efficient gene transfer into mammalian and nonmammalian cells. *Proc. Natl. Acad. Sci. USA* **90**, 8033-8037.

Chen, W., Burgess, S., Golling, G., Amsterdam, A. and Hopkins, N. (2002). High-throughput selection of retrovirus producer cell lines leads to markedly improved efficiency of germ line-transmissible insertions in zebra fish. *J. Virol.* **765**, 2192-2198.

Concordet, J. P., Lewis, K. E., Moore, J. W., Goodrich, L. V., Johnson, R. L., Scott, M. P. and Ingham, P. W. (1996). Spatial regulation of a zebrafish patched homologue reflects the roles of sonic hedgehog and protein kinase A in neural tube and somite patterning. *Development* **122**, 2835-2846.

De Martino, S., Yan, Y.-L., Jowett, T., Postlethwait, J. H., Varga, Z., Ashworth, A. and Austin, C. A. (2000). Expression of sox11 gene duplicates in zebrafish suggests the recriprocal loss of ancestral gene expression patterns in development. *Dev. Dyn.* **217**, 279-292.

Dermitzakis, E. T., Reymond, A. and Antonarakis, S. E. (2005). Conserved non-genic sequences–an unexpected feature of mammalian genomes. *Nat. Rev. Genet* **6**, 151-157.

Frankel, W., Potter, T. A., Rosenberg, N., Lenz, J. and Rajan, T. V. (1985). Retroviral insertional mutagenesis of a target allele in a heterozygous murine cell line. *Proc. Natl. Acad. Sci. USA* **82**, 6600-6604.

Gaiano, N., Allende, M., Amsterdam, A., Kawakami, K. and Hopkins, N. (1996a). Highly efficient germ-line transmission of proviral insertions in zebrafish. *Proc. Natl. Acad. Sci. USA* **93**, 7777-7782.

Gaiano, N., Amsterdam, A., Kawakami, K., Allende, M., Becker, T. and Hopkins, N. (1996b). Insertional mutagenesis and rapid cloning of essential genes in zebrafish. *Nature* **383**, 829-832.

Grabher, C., Henrich, T., Sasado, T., Arenz, A., Furutani-Seiki, M. and Wittbrodt, J. (2003). Transposon-mediated enhancer trapping in medaka. *Gene* **322**, 57-66.

Hauptmann, G., Söll, I. and Gerster, T. (2002). The early embryonic zebrafish forebrain is subdivided into molecularly distinct transverse and longitudinal domains. *Brain Res. Bull.* **57**, 371-375.

Houart, C., Caneparo, L., Heisenberg, C. P., Barth, K. A., Take-uchi, M. and Wilson, S. W. (2002). Establishment of the telencephalon during gastrulation by local antagonism of Wnt signaling. *Neuron* **35**, 255-265.

Jahner, D. and Jaenisch, R. (1985). Retrovirus-induced de novo methylation of flanking host sequences correlates with gene inactivity. *Nature* **315**, 594-597.

Jahner, D., Stuhlmann, H., Stewart, C. L., Harbers, K., Lohler, J., Simon, I. and Jaenisch, R. (1982). De novo methylation and expression of retroviral genomes during mouse embryogenesis. *Nature* **298**, 623-628.

Jowett, T. and Lettice, L. (1994). Whole-mount in situ hybridizations on zebrafish embryos using a mixture of digoxigenin- and fluorescein-labelled probes. *Trends Genet.* **10**, 73-74.

Korn, R., Schoor, M., Neuhaus, H., Henseling, U., Soininen, R., Zachgo, J. and Gossler, A. (1992). Enhancer trap integrations in mouse embryonic stem cells give rise to staining patterns in chimaeric embryos with a high frequency and detect endogenous genes. *Mech. Dev.* **39**, 95-109.

Lin, S., Gaiano, N., Culp, P., Burns, J. C., Friedmann, T., Yee, J. K. and Hopkins, N. (1994). Integration and germ-line transmission of a pseudotyped retroviral vector in zebrafish. *Science* **265**, 666-669.

Linney, E., Hardison, N. L., Lonze, B. E., Lyons, S. and DiNapoli, L. (1999). Transgene expression in zebrafish: A comparison of retroviral-vector and DNA-injection approaches. *Dev. Biol.* **213**, 207-216.

Meng, A., Tang, H., Ong, B. A., Farrell, M. J. and Lin, S. (1997). Promoter analysis in living zebrafish embryos identifies a cis-acting motif required for neuronal expression of GATA-2. *Proc. Natl. Acad. Sci. USA* **94**, 6267-6272.

Miyashita, T., Yeo, S. Y., Hirate, Y., Segawa, H., Wada, H., Little, M. H., Yamada, T., Takahashi, N. and Okamoto, H. (2004). PlexinA4 is necessary as a downstream target of Islet2 to mediate Slit signaling for promotion of sensory axon branching. *Development* **131**, 3705-3715.

Naviaux, R. K., Costanzi, E., Haas, M. and Verma, I. M. (1996). The pCL vector system: rapid production of helper-free, high-titer, recombinant retroviruses. *J. Virol.* **70**, 5701-5705.

Nobrega, M. A., Ovcharenko, I., Afzal, V. and Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science* **302**, 413.

Nornes, S., Clarkson, M., Mikkola, I., Pedersen, M., Bardsley, A., Martinez, J. P., Krauss, S. and Johansen, T. (1998). Zebrafish contains two pax6 genes involved in eye development. *Mech. Dev.* **77**, 185-196.

O'Kane, C. J. and Gehring, W. J. (1987). Detection in situ of genomic regulatory elements in Drosophila. *Proc. Natl. Acad. Sci. USA* **84**, 9123-9127.

Parinov, S., Kondrichin, I., Korzh, V. and Emelyanov, A. (2004). Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes in vivo. *Dev. Dyn.* **231**, 449-459.

Pfeifer, A. and Verma, I. M. (2001). Gene therapy: promises and problems. *Annu. Rev. Genom. Hum. G* **2**, 177-211.

Prince, V. E., Joly, L., Ekker, M. and Ho, R. K. (1998). Zebrafish hox genes: genomic organization and modified colinear expression patterns in the trunk. *Development* **125**, 407-420.

Rimini, R., Beltrame, M., Argenton, F., Szymczak, D., Cotelli, F. and Bianchi, M. (1999). Expression patterns of zebrafish sox11A, sox11B and sox21. *Mech. Dev.* **89**, 167-171.

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* **21**, 99.

Sanes, J. R. (1989). Analysing cell lineage with a recombinant retrovirus. *Trends Neurosci.* **12**, 21-28.

Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D. and Gaul, U. (2004). Transcriptional Control in the Segmentation Gene Network of *Drosophila*. *PLoS Biol*. **29**, e271.

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.

Spradling, A. C., Stern, D., Beaton, A., Rhem, E. J., Laverty, T., Mozden, N., Misra, S. and Rubin, G. M. (1999). The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. *Genetics* **153**, 135-177.

Thisse, B., Pflumio, S., Fürthauer, M., Loppin, B., Heyer, V., Degrave, A., Woehl, R., Lux, A., Steffan, T., Charbonnier, X.Q. and Thisse, C. (2001). Expression of the zebrafish genome during embryogenesis. ZFIN Direct Data Submission Unpublished.

Thisse, C., Thisse, B., Schilling, T. F. and Postlethwait, J. H. (1993). Structure of the zebrafish snail1 gene and its expression in wild-type, spadetail and no tail mutant embryos. *Development* **119**, 1203-1215.

Wilson, C., Pearson, R. K., Bellen, H. J., O'Kane, C. J., Grossniklaus, U. and Gehring, W. J. (1989). P-element-mediated enhancer detection: an efficient method for isolating and characterizing developmentally regulated genes in Drosophila. *Genes Dev.* **3**, 1301-1313.

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.

Wu, X., Li, Y., Crise, B. and Burgess, S. M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**, 1749-1751.

Table S1. Flanking genomic sequences of transcriptionally activated proviruses	
CLGY number	Flanking sequence, 5′ to 3′ direction
1	TAAAGGAGTAATCCTCACTTATTCTCTTTAGTTATGAGTCATGTGGTCGTCGACGGTTCACACGGTCTAGATCAGCAGG
2	TAAAAGTGGAATTTACATATAANAATAATAATAATAATAATAATAATAATAATAATAATAATTATTATTATTATTATTAT
3	ATTACATAGTTCATTCATACATTTCAGTTTAGCTGCGCTTGTCATTCTCAGAATCTGCGCAGATTTTGCTGAACATATGACCTAATATATAAGCTGATTATGGTGAAATTTAC CATTATCATTTTACAGGTGTTTTACCTGTGTTTTGAATTATGTATTGTAGAAAATGGGATGGTAATTTATATTTATTTATTGATTTACATTTTCTCTCTTTGCATCACTTTCT GGTTGACAGGTCTAGCACCTGTCGTTTTTCCTCAATGAGGTGTCTTGTTAGTATATGATAGTTACGTAATGAATTACAATGCAGGTATGTGATTGGTATTTCATATATAATGA TGAAAGGTGTGATTGATAAATAGAAGTATTTAGGAATGATCACATCGTCACTGAACAGTTGACATAAAATAGACTTTTATTATTTGTTATAAAACATTTGCAGATGTTGTCCA TCTTCGTTTAGTGCTTTCCATTGTTGCTTGATTGTAAGACTTGTTT
4	AATAGTCTTATTAGTAATTTTTGCTACACTCCCATTTCAAAAAGTCCCTTTTTTCACTACATTGTTTTTCAGTTGTTTGGACTCCTTGCCCTCAACTGTCCACACTATATTTT ATAGGGCCCAAGAGATTGCAATTTAGTAAATAAAGGACGAGGACGAAAATAATAATAACATTTATTACTTGATAAATGCTTCTGTTGGGGCAGCACGGTGGCACAGTGGGTAG CACAGTCGGCTCGCAGCAAGAAGGTTGCTAGTTTGAGCCCCGGCTGGGTCAGTTGGCATTTCTGTGTATAGTTTGCATGTTCGCCCTATGTTGGCGTAGGTTTCCTCTTGGTG CTCCGGTTTCGCCCTCAGTCCAAAATCATGTGGTACAGGTGAATTGGGCAAGCTAAATTGTCCGTCGTGTATGTGTAGGAATGAGTGTGTATGGATGTTTCCCAGTAATGCTG AGTAAAATATATGCTGG
5	TTGTTCTAAAATGAAGTATTCTGAAATGGAAGGAGTTTTGACAGTTGATTAGTTTTTA
6	TAACCGGGTTAAGCCTTCAAATTGGACTTGAATACTAGTAAGCTGAATACTAGTATCTTGAAAAATATTTAGTCAAATATTATGTTCTATCATCATGGCAAAGATAAAAGAAA TCAGTTATTAGAAATGAGTTATTAAAACTAAAGAAATGTGTTGAGAAAATCTTTTTGTAAACAGAAATTAGGGAAAAGTGCAGGGGGCTAACAATTCTGACTTCATACACTCT CAGAAATAAAGGTAC
7	AGGATGCTTTGAAACTTTTCAAAACTCCAAACTAGCAGATTTTTCCTCATTTACTTCTAT
8	ACAGTATATACAAATAAATATGTATATACTAAAACACAGTTTGATAAATACTTT
9	ATAGTACACAACAGANTGAACCGCCAACTTATCCTGCATATGTTTTACNCAGCAGATGCCCTTCCAGCTGCAACCCAGTACTGGAAAACACCCATACACTCTCACAGTCACAA TCATACCTTATGGCCAATTTAGCCTATTCAATTCACATGTAGCACATGTCTTAGGATTGTGGGAAAACCAGAGCACCCGGAGAAAACCCACGCAAACAAGGGGAGAACATGCA AACTCCACACAGAAATGCCAACTGACACAGCCGTGGCTCAAACCAGTGACCTTCTTGCTTTGAGGTGACAATGTTA
10	AAATTTGACAAATACCATCAATTGTGTTCAAGCCCCCCCACACGCACATAAACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC ACACATACACACACAAACACACACATTATTATAGATATACCTTCATAATTATGGTTA
11	AAAACCAAAAGAGTATAGTTCC GTTCCTAGACATATCGGTCTAGAAAACAGCAAGTTTGAATTTTATCATCGAGCTTAGTACATGGTATAACCACAATCTAGCTTT
12	TAAAATAAATGTCACAATTTCAAGATGTAAACTTAGAAATCTGAGAAATAAAGTCGGAATTGCAC
14	TAAAAAAATAAGAAAAAAAATCGAAAGGGGCTAATAATTCTGACTTCAACTGTATATATATATATATATATATATATAT
16	TANTGTTGAATATTAGAGCAGTTTTTTTTTGCATTTAGTTGCCAACTACACTTGAGTGTATGTAAGATGATATGTGTCTTTATGCAGCAACAACTGCAGCATTGTTGTCCATTA GCATTTGTTTTGTTTTTCCATCTCTTTTTTTCTTTGTATGTTGCTTTTTTATCATCTTGAATACCTTGATATGTTTTCATATTCATAGTTGTCAACGATATGGATTTTCACTC TGAAAGACAAGAGGTGCATTTGGGTAATCTACACAAAAGTATGTTTATA
18	AACTTATTGTATCCTATTCATCAAAACCTGTAAATGTTGCTTGGATCTCATAAATATAAGCTAACTGTCACCATATTCATATAGAAACCATTTTTGATAAATTTTGAAATATG TTGCATGTACATTTGACATATTTAATTCCTCCACGTACCAAAGTTTGAGAACCACTGGACTAGATATTTTGCAATATACTAGTATTCAGCTTAAAGAGACATTTAAAGGCTTA ACTAGGTTAATTAGGTTAACTAGGCAGGTTAGGGTAATTAGGCAAGTCATTATTT
19	AAATAACTAAAATATCTTCAGTAATGTCTACTTCTGCTGACCGTATTCCCATTTATTTGAACAGATGCAGTGTGTTTTGTTTCTGACTAACTATGAAAACAAACCCGAGAAAT TCCTGTGCTTCATGACTTCAACTTGCAGAAACTCCCACCAGGCAGGCAGGCAGACCCAAAAGATATTTGAACAGGAGAGGCAAGCGTCTGGGAGAGAACAGGAGGCCAGAATC CCCAAACGGCAGCCATTGTGAAATTTCCATTCTTATCTCAATGACAGCGGCTGCGCTGCCGAGGTGGGAATCCCTCTATTTACACTCTACTTCCTCTCTCAGCAGGGAACTGT GGGTACTCATTACCCTCACAAGAAACCCTGTGCATTAGTGCACTCCCTCTGACTTCTGCCTCACAAACAATATTTCACTTCTTGTCTTGCTAGTGGCCGGTAGCTAAGTCTGT GTCAGTCATAGTCACGCATAAAAAGTTTACAATATATGATCAAGTATTTTCATTTCACAAACTAGCCTGTGAGTTTCTGAAAAAAAATATCTTTAGTA
20	TAGAAACTACATAAACATAAACTGTTTAGATATTGA
21	AAAACAAACAAAGACAAAACTTACCTTTCACAGATCAACTACAACTTCACTGAGCAGTAGTCATGGCACATCATGTTCT
22	TAATTTAGTAAATTTTCATATTTATTTTTTTACTAGTACTTACAAACAAACATTGTCCAGTCTGACTTTGGTAATCACCAGTCTTGCAAGAAACAAAAGTATGCCAGACGTTA CATAACGATATGTTATAT
24	GTATGATAATGGATAATAATGCNAATATATTCAAATATGTTTATCCAATAGAAAAATACATAAATAAACACGCCTGTGCTGTGTAGAATAAAAAGAAATTAGCTACTGAGTAA AAACTATTTA
30	TAACTGTTCAAACTTTTACAGTATTTTACGATGACTGGAAAATGAAGTATGTTACACACAAATTCTGGAACATATCTACTTTGATATTATCTATAAACCAATAATGTACTGTA ATGTTGAAAAGCATTACATCTGAAAAAACATCACTGTTTACACAAAACCAAGAATTACAGTAAAAAGACAACAAGCATTATCACAATAATTTGACTAT
31	ACTCTAATAGTAATTATTAGTAGTAACGTTAGTGCATATAGTGTATGTATAACGTTATTACGTTACCATTTACTACCTTTT
32	ACACAAAGATTTAGCAAAAAAAAAAAAATCACTATTCAGCTTACAAATAATTCTTATAAACATTTTTGAAGTTATTGTTATGTACTCTTCTTAGTGTGGTTTT
33	TAACGGTCTAATTGCAGACGGGGGAGAGAGAGAGAGTGAGAGAGAGAGAATAGGGAGGGACTTTGAGAGACAAAAAAAGTGACACTGATAAGGTTAGCAAGAATGGAGTGAGG GAAAACAAGTTAGCCTGTGTATACAGGAGGGGAGCAAGGGGGC
35	AAAAAAGAAAAAAAATCAAAAGGGGGCTAATAATTCTGACTTCAACTGTATGTGTTGTAAATCTATTATGCTGCAGAACAAAATCACCACAAAAAGGGGAC
38	AATAATCCCAAGGATTGGACCCGATTTACATCTAAACAGTTAGAGGTACTTGCCTTTACCTCCGTATTGGCTAGATGACAAATAACATTTTTTAGGAAGTACTACTAAATCAC CAGGTATTTCATGGTGGCTTGT
39	ACCTCTGCAGGTCTTTACATGCCAACATCCAGTTACTCAACATATCCTGTTGTCCACGTATGCTACATATATTT

41	TCATCATTTTGGCTAAACGTTCACGTGGATTGCAATTTGAGCAATT
47	TAATATATAACTATAGGTTATTTTGCAT
50	ACATATAATCAAAGTAAAGGATAAACTTATTTATTATCCTATTCCTTATTATTATCCTAATAATTAGCATAATA TTGCATAATAATT
65	AATTTTGACACCCACGGACATCGTAACCAAATATAAATCAGAGACTTTCGCATACATAGCCTCATTTCAGTTGTCAAGTTTTGTAAAAACTTGATTTGGTGCGGATCAGAGTG CGATTGCAGGTTTCACATATGCCAGACGAACCACGCT
68	GTATTGCAGAAGTCGTTTGACATGTTACGATGCCTCAATTTT
69	ATTAGTATATTGTATATCAGCGTTGCATATTTCAGACGCTAATTAGCTAATAATGTAGCTTTAGACCGAGGATTTT
75	TAAAGGTATTGAAATGGTAGGCAGTGTGTTGCTTAGGCACAGTGTTGACGCTGTGCTCTGGTTTGTTTGTTGGGTCTGTTGTGTGGTTTTGGCTCTGTCTG
77	AGTCGCAAATCTATTTAGATAACACTGGTTCAGTCGGTGTTCACAAAAACATTATTTTTCAGATTTCTCAAGCTAGTGCCCGTTCATAACATAATTCTGCACATCTACACGAA TATGAGCTATCATCAAGATGGACAGATAGCTAGCGCTGGGAATACGATGATTGTTCAGGTTTGATCTCCGTGAACAACGGCCATGTCTGAACCCAGAAGAAGCGGCCAATCCT GTGCCGTTTCTTCTGCTGCACCCACGTGGGATCCCTTTGCGGCTTTACAGCTTCTCAGCCAATCGCAGAGCAGCATCTATATTCTCTCGACTGCAATTGGACGGGAAAGGT
86	TAGATTACAGAATTTTATCGTAAAATAACAGATGGTAAATTACAGTAATTTACCGG
97	AACTTATTTACTTTTCTGCATGGTTTCCTTGATCACAAACTCTAATTTTATTGTTTTACCAGTTTGGAAAATGCCTGCACTGTTATTGAGCGAATGAAAAATAAGCGCAGTTG TCGTTAC
101	AAAAGTGGAAGATCAACATCCCATAAAGCAATTCACAACCGTCAATAAAAGGAAAACACTTGTAATTCACAAATTATGGAAGCTGT
102	TATCTGACCTGTTGTTTATTTATTTGTTGTAATAAACAACAAATAAGCCTGTAACAAACAATGTCACGCATTTCGTTACTTTTGCAATGTAACTTTTTCATCGGTCAAACTTG AGGTT
104	TAAAGGATTGATTGCAATTGTTT
109	AAAATTTACATTTCATGTAAAAGACAAGAAAAAAAGAGAGAAAACTTCTGTTTACACCACAATGTGTTTTTTATTTGTGAACTGCCCCTATATCTGCATACCCTAC
128	AGAAAAGTTTTTTGTTCAACAGGAAGATTTTTTTTTTACCTTGAACTTTTTGTGCATCTTGTAAGTT
129	AATATTTTTCTATATCATATACATTTGGGTGTACTAGTTTTTAGACCGTTAGCATAAGTTCTTTTGTTAAATAAGCTCCAGATTTGGCTTCAGTACTGGCTAATCTAATGTAT ATGCACAAATATAATATTGTATAGCTTCCTTTT
130	AAATCATTTTATTCAGTGTACATAAAATTGTTTCCCGTTTTCCAGATGACATTGTTAGTTGTTGGACAATATCGCGCTCTTAGTAATGCGTTACCGG
131	TAGTTTCAACTTCAATTTTGTCCAGAAAGATTTTATTATTTGCCCCCAACTGCTCTACAATTATAAAAGTTTACATTATTTACTATTTTTATGTTAGTCACTAGAAATAAAAG ATTCCCATATGCTCT
135	TAACCCAGGTCAGGGCAAGGTCGTGACTTTGGCAACTTACCCCTAGACCGTCTGCGCTGTCTGCTCATCCATAAACGCATAAACGCTCAATCAGACTGACTTTGAAATACACT TGAGAAACCTGCCAGGGGACAAAAGC
142	CTTGTAGAAATGATACTTGTAAATAAAGATTCATGTTTGTACTAAATGTTA
143	TGCTAACCACTGGGCCGCCGTGCTGCCCAATCACAGGTAAAGGAGGAGAATAGGGGAGGAGGGGGGTTTCTTCCAAACGAAGATAAAGTGAACTGAAAACTGTGGCTATTTAT AGTAGCTTAGGGCTTATATGATTGGAAGATAGTGATTAGCAAATGCGAGACTGGCCGTGATAAATCATAAGCACGTGATCCTCTCGAAATTAGTTTATGAATAAACTTCACTT TGGAAATATT
151	TAAAATAGCAATTTCTAATGT
154	TTATGAATGTTGCTGACGTTAGTAAAACAAATCATTGTTCATTGTGCGTTACTTCTA
155	TAATGGAGGAAGTTAACTTTGAGCTTTGTAACAAACTAATGATTTATATATATATATAA
158	AGCACCATGTGGTTGATCATTGCTTCTCCAGAATTACACTGTGGATTCCCTGTGATCCAAAAATATTT
159	TTATAATATAGAAATATAAAAAAAAAATAAAAATAGATT
160	TAAAATGAGCAGACTCAATGCAGATCCTGACTATTTTTGAGGGACAACTTCATTGTTT
162	AAAAGTCAGAAACTCTACACAGCTGCACACAAATGGTTGTTCAGAAGGTCTATGTCAATTTTGGTTTAGAGAATCACTAAATATGCC
182	TAGCGGAGAGAAGATTTTTACCACACATTTCTAAATATAATAGTTTTAATAACTCATTTCTAATAACTGATTGATTTTATCTTTACCATGATGACTGTAAATAATATTTTACT AGATATTTTTCAAGACAC
183	TTCTCAAAGACAGAAAAACAAACATACTGTTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGCGTGTGTAGAGGTAATGTATTGGACATGTCTCTCCATCA CTGGTTTTGTTATTTTATGCAAAANAAATGCTGTAAATTACATTATTTACATATTTTTTGTTTACATTTGGAATAAATATAGTGTACAATGAAAAAATGATGTCTGCATTATT GTTGCGCAACAATTTATATGTTTA
185	AATAAAAGTAGACGCTTTCGTAAGTAAAAGTAGCTACTTTGGTATTGCATTTCTTTTCAATTATATATGCAAGATGAGCTACTTTTGGCCCTTTAAACGCAACAGGAAAAGTT TGACTGCTCTTGTGCTGTCAGATATCACATGAGGTTTAAGTAACATAAATAGATGAACTATATT
186	TAAACAGTAAAAAGCTGTAAAAATGCTTACCTACACTGTAGACATGACCTTAGTTTAGTACAATTTTATTATATTATTTTATTTTTTTAGTGTAAGGTAGCCAACATGATTTA TTATTTTCAGATTTATCTACTCTGTAAAAAATACTGTAAA
188	TAAAGTGTGGTAACCATAGACTGTAAAATATAGGTAACCATGTTTTAGCTCAATGCAAAGTGTATAAGCCTTACTGTTGCAATAATAATAAATAAAAAT
189	TAGTTGTGTTTACATGCCCCATGTGGTGCACACTGCAACATGACATTATGACAGCTACATATGTCAACTGAAATAATGAAAGAAAAGAAAAAAATCTGGATCACAGCACAACA CACTTTACAGCATAGTTCAAAACAATGTGTGCACAAACCTGGAACTACAGCCGACATTTAGAAAGAAAAATGATGTCAC
195	ATACGATGATCTCCGTGCTGCCGAATTGGGGGTGGGGTTGCGTTT
196	GTATAAAGTCAATCTTTTTGTTTATTTATCATATTTATCATTTTTATTTACCACAAAAAATGAGACAAATAGCATAAAAACTGCTGAAAAAAATAAATAGCAACACTA
197	TAATGTGGTTTGATTTTGATCAAGTGTTTGCATTTGCAGAGAGCTACAGTCTGCACTGCATGATATCAAGTGCCTCGGCTGAAGTCATAGAAGCAAAGTCATGAATGAAT
198	AAAACGAAATTCGTTATTATGTTAC
202	TANCAGAAATAATGTACAGTATAGAATATAAAGATTTGGTGCAGTGGAAAAAGAATGAATATTGCATATGACTCCCATCCATGAGTCTGGAGGACTGCATCCATACATCTCTG CAATGAATCAAATAATTTATT

205	CTAGGTAATTATTTGTGCAGCATCTCACTTCGTAGAAAAANT
215	AAACAAAATAGTTTTGTTTCTATGCACGTTTGTGTATACAAGTGTTTCCTCTAGGATTTTTTCCAGCTGTGGCAGCAGACTCTGTTGGGTCTTTTGCACAGATCTACCAACTA CCTGTGGTGTTATTTCAATGACAAATGTCATGAGCGCAGTATTACAAGTCGAGATCACATTTATGTTACAGAAGCATGCGAATCTCTTTGCTTGCACGCCGATTTTCTCTGTT CGCGAACAAAGCTTCTTTCATGCCCTCATATACACTTCTCAAGTGCAGACTTTCTTGTGCGCTATCAAATAAATACTGATGAAGTGCGATTTAGTGCGTTTATATAACAAGCA TGTCTCCAACATTTGTTATATTTGCTAGGAATATTTACGAATGTCTCCAATAGACCTACAGAGCGGCATTAGCATCCTAAAGTAATACAGCCATAAACTCCAGCAAGATAAAG TCATCTATAAATTTGAAAGTATTATAAAGCATTATCTATTATCTATAAAGTATAATTATGGAAACATTGTTCATCCT
229	AAAACAAACGCCAGATTCAGAAAGATCACATTGTTCACAGTCAACT
232	TAACTTGTGCGTGCTTTTCAGGTTGCTTCTTTGTAACT
236	GTATCATGTATGTGTACAGTACCATTTGTAAGCACTATTTGACATTATATTATAATTTATTTTTCTGAACCCTCAATTCAATGTCCACACTACAACATGGAAATTATTTATTT CACTATTGGTACATATTATTTCTCAGTGTGGATAATAGACTGAAAAAATAAATAAAAGTATTTAGATTCATGTGAATGTACCCTAATATTCAGA
237	GATACAATGCGGATATTTCTCACAACGTGCTAAAACGTTTGGTGTCTGGGAACTGCGCGCACCTAAGAAAACACGGCTGATCGCTGTAGTTTGTATTGCTCACAGCTGAAGAA GACATAATCCTGCTTAGCTGTCCTTCAGATCAAAGACAATTGACTGTT
238	TAAATTAGCTTCATTGTATTGCAACATCCTCTGACAAAGGACAC
239	TTCTTAGAATATGAATGATTATGGTGACAAATGTTTTCAGCGGTCTTT
258	GGGTTATATATAAGTTTGTGTGTGTGTGTGTGTACATTTTTTTATTTA
263	TAGATGCAACGAGTTGCGCACGATATATTAGTACATACATAAACAGTGTGCGTGTCTTGCATGTTCAACTTTGCATTATGAAGTGAAACATTCATTCTGTCAATAGCTTTCTG AAAAATATGGCGAATAATATTTTACCAGTCAAATAAAGCCCCCAGTGTAACGTTAGCACTTTTTTGCAATCATTTTTGTGGTAGCAGTGACCTACAATCTCACTTCATAA
264	ATCAAAAAGGACTTTCACGGTAATCGCCTAATTTACTGCTTAACTTATTCAATCAATGCAGTGTACAGCCCCTTGCTGTCTAAAAATTTTATAAAATTAGCCAGGGTGGGTGG TTCGAAAGCCCCACTGACAAAGTTCCTACATGGGCTCACTTGTCCC
266	AACCTAGTAGAAGAAAATCACGTTTTCTGCATATCTTAGAAACCACCTAAGACACCCTAACAACAACCTACCCATATCCTACACTCAAAAGCTGACTTTTGCTGCTTGTTCAA ACTACTTATTTAAAATTAGTTGAAACAACACAATTCTTAAGTGTTAATTGGGACAAATTTAGTTTTATGCTCAATCCATTTGAATTTGTAACAATAAAGCTA
268	AAAAAAAATTATCAAAACTATTTATTTATTTTACAAAACAACTGTATTTACATTGCTTCATTTTGTTTTACTCCATTTTTACAATTACTAAATTCATGTTTTGGAGAATGTAT TTGTTATATATATATATATATATATATATATATATATATATAT
287	AAAAAATGATGAAAATGTAATTTTGATCAAATATGATTAT
298	TAACATGTGCGATTTTGCCATTGCACTAAGCCAGGGGTGTCCAAACTCGGTCC TGGAGGGCCGGTGTCCTGCATACTTTACCTCCAACCTCCCTCAACGCCTCCCTG
299	AACTAAAATGAATTAGTTTACTCAGTGCTGTTGCAGAAATAGTTTCAATTATGTCTGATTTACAATTATGCATCACATG
308	AATCAGGAAAACTTTTCAGAGAGGTTACAATTGAAAATCTGTGAGATAGATGGATAACACACTA
318	TAATGTTTATTTACCAGATTTGTTTTTCAATTTTTACTCCCAGCATTCGAGATTGATGGAAAGGAGTTGTTTGGCTCACTTCAGCCATTTTGTGTTTATGGTTATAAGTTTAT AGACCCTAAAGTAATTTAGCTTTTTACCATATCAGTTTTTGATAAAGGTAGTGTGAAATAAGACTCTTATCAGTGAGTTTATTTTCCATTTATGATCTGTACACAAACATACC TCACATTTCAGGTTTGTAAA
368	GTGATACTGCCTCATTTTTGACTAAATTTTTCACACTCTGAGAATTTAGTTTATAGTCCTTGCTACTGCTTAGATTTGTTATCTAGTGATTCACAACCTAATAAACTGTATTG ATTAACAGTAATGTTCTTATGGCAGTGCCACAGAGGAGTCTTATGACTAGGCCAAGACAGAATCTGCGCACAAAGAAATCCACTGATTTCTGCAGATTTTTAGCCCATCGTTG AATCTATGTTTTTATCTGCGTAAATGTGTGTCAATCTATATTTATTCAGTTTTTTTAATCATCTCACTAATTTTAATGTGATATACTA
369	TAAAACCTACGGACGCTATCTTAGATAATCAATACCCGACTTTGGGCTGTAACATAACAGTTTGTTCCTTGTTCGGTCCGTGTTTGCATAGTGTGAGAAAAAAACCTTTCTTC CTGTAGAGGCATAAAGAGATGCATAGTCAT
372	ATAACTACTTTATTGTTTGTATTATTTTATTATTTGAAATATTT
373	TAATCTGGGTTCTCATTCACACACATACACTACGGGCAATTTAGCTTACCCAATTCACCTGTACCACATGTCTTTGGACTTATGAGGGAAACCAGAGCACTCAAAGGAAACCC ACACGAACATGGGGAGAACATGTAAACTCCAAAGACATGCCAACTGACCCAGCCGAGGCTCGAACCAGAGACCTTTTTGCTGTAAGGCGACAGTAC
375	AAAGCAGAAGAAAGTGGCGAGCGGATTCCGCCAGTCTACCTCTGATCCTGTTGCAGCAACATTTGGAGAGGTAGCCCCTTTGCAAACCCTGAAAAGGGTGGTTT
376	TAGAAAAAAAAAAAAGGTGATAAAAAACATGACAGAGGAGCTAAATTTATTATTTGTAAAACTACTAAAGAGTCACAAACAGCAAACTTCAGTTAGTGTTTTTTGAGTTTTGT TTTTCGAAATATGAGGTAAATGATAC
379	AAGGCAACCAGGATTTTTTACAGTGTACTTATGCTCAAGGATTTTTTTATTTAGGCATGACTTTATGGTTGGCATACTGAGCTAATTAGCATAACCAGTGTTTTA
381	TAATTTTAAAGCAGCTATCATGATATTATTTATCAATATGTGGCTCTTTTTGAACTCTCAGAAGCTCTAAAAATTAAAAATGTAAAACAGGAAATACGTTGGGAC
383	CAGCAAACTGTGAGATATGTAAACTTAGGATTGCCTGCTGGGTATTTTAAGTTGAAGTGCTATTTTTATATTTAATTTTTTTATATTTATTGTTGCACTAAAGTCCAAAAGTG AAGAATTTGTGTTATTTCTTACTTGACTGTTCAAGCTACCTCACAGAAGTGCTCGGTTTGTTAATAGAGATTTAACATTTT
391	TAATACATTTGATCACAAAACCAGCGATTACAAAAACTGTGAGAGTCATCTACAAAAAAAATGTAGCATCAAACTGTTTTTTATGGTGTTGTTG
393	AATTTTATTTTATGGCCTAAAAATCAGTGTCAATCGCAGACATGCGCAATCAATAGGTTGTTTGCAGTCACGTGGTACTAATGACACGAGATGGCGCTCAGGAAGTAACGTTA TAAAACTGAAAGGTAACTTCTTATTCCTGGGAAACCTTTTTTTATCGATTGTTGGTTTGAAAGAGTTTTTGTCT

Flanking sequences were blasted against the Ensembl zebrafish genome browser (zebrafish whole genome shotgun assembly sequence version 4) (http://www.ensembl.org/) to obtain genomic location of activated insertions. In some cases, it was impossible to obtain a match using only the closest flanking sequence. In such cases, the flanking sequence was blasted against the zebrafish whole genome shotgun (wgs) trace sequences at NCBI (http://www.ncbi.nlm.nih.gov/) and Ensembl, and wgs sequences overlapping with the proviral flanking sequence were used to build a larger contig.