Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

A large-scale zebrafish gene knockout resource for the genome-wide study of gene function

Gaurav K Varshney, Jing Lu, Derek Gildea, et al.

Genome Res. published online February 4, 2013
Access the most recent version at doi:10.1101/gr.151464.112

Supplemental Material	**http://genome.cshlp.org/content/suppl/2013/02/04/gr.151464.112.DC1.html**
P<P	Published online February 4, 2013 in advance of the print journal.
Accepted Preprint	Peer-reviewed and accepted for publication but not copyedited or typeset; preprint is likely to differ from the final, published version.
Creative Commons License	This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2013, Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

A large-scale zebrafish gene knockout resource for the genome-wide study of gene function

Gaurav K. Varshney*[1], Jing Lu*[2,3], Derek E. Gildea[4], Haigen Huang[3], Wuhong Pei[1], Zhongan Yang[3], Sunny C. Huang[1], David Schoenfeld[3], Nam H. Pho[1], David Casero[3], Takashi Hirase[3], Deborah Mosbrook-Davis[1], Suiyuan Zhang[4], Li-En Jao[5], Bo Zhang[6], Ian G. Woods[7], Steven Zimmerman[7], Alexander F. Schier[7], Tyra G. Wolfsberg[4], Matteo Pellegrini[3], Shawn M. Burgess^[1], and Shuo Lin^[2,3]

1. Developmental Genomics Section, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
2. Laboratory of Chemical Genomics, School of Chemical Biology and Biotechnology, Shenzhen Graduate School of Peking University, Shenzhen, China
3. Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles, Los Angeles, USA
4. Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
5. Dept. of Cell and Developmental Biology, Vanderbilt University, Nashville, TN, USA
6. Key Laboratory of Cell Proliferation and Differentiation of Ministry of Education, Center of Developmental Biology and Genetics, College of Life Sciences, Peking University, Beijing, China
7. Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

^ Corresponding Authors: shuolin@ucla.edu; burgess@mail.nih.gov
* These authors contributed equally to this work.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

With the completion of zebrafish genome sequencing project, it becomes possible to analyze the function of zebrafish genes in a systematic way. The first step in such an analysis is to inactivate each protein-coding gene by targeted or random mutation. Here we describe a streamlined pipeline using proviral insertions coupled with high-throughput sequencing and mapping technologies to widely mutagenize genes in the zebrafish genome. We also report the first 6,144 mutagenized and archived F_1's predicted to carry up to 3,776 mutations in annotated genes. Using *in vitro* fertilization, we have rescued and characterized roughly 0.5% of the predicted mutations, showing mutation efficacy and a variety of phenotypes relevant to both developmental processes and human genetic diseases. Mutagenized fish lines are being made freely available to the public through the Zebrafish International Resource Center. These fish lines establish an important milestone for zebrafish genetics research and should greatly facilitate systematic functional studies of the vertebrate genome.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

The zebrafish genome, along with the mouse and human genomes, are the only three vertebrate genomes that have been sequenced to a degree that they can be considered "finished" (Waterston et al. 2002; International Human Genome Sequencing Consortium 2004). Gene knockout remains the fundamental mechanism for deciphering protein function *in vivo,* and the first step in leveraging the full power of a model organism's genome project is a systematic mutation of all genes. In the last two decades, zebrafish has rapidly become a widely utilized model organism for studying vertebrate development and modeling human diseases. One of the primary reasons for the popularity of zebrafish is that they are particularly amenable to genetic studies, allowing the identification of mutations affecting both embryonic development and adult homeostasis. For zebrafish, most "forward" genetic studies have been conducted using the chemical mutagen ethylnitrosourea (ENU) (Solnica-Krezel et al. 1994) followed by screening for phenotypes of interest and positional cloning of the mutated genes (Talbot and Schier 1999; Bahary et al. 2004). However, for a systematic approach (*e.g*. one that allows testing gene function for entire classes of genes, or even the entire genome in a non-redundant fashion), it is more effective to first create mutations in all genes and subsequently evaluate the effects of these mutations (*i.e.* "reverse" genetics). Because zebrafish are amenable to large-scale screening efforts (Mullins et al. 1994; Amsterdam and Hopkins 1999) and they now have a completely sequenced genome (Howe et al, *in press*), they are an ideal organism for systematic reverse genetics in a vertebrate, and testing all protein coding genes in the zebrafish genome via reverse genetics is an achievable goal. As an alternative to ENU, Moloney murine leukemia virus ([M]MLV) based insertional mutagenesis has been demonstrated to be an efficient approach for

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

mutagenizing thousands of genes both in mouse embryonic stem cells (Friedel et al. 2005) and in large-scale zebrafish genetic screens (Amsterdam and Hopkins 1999). A major advantage of retroviral mutagenesis over ENU is that it allows for rapid identification of the mutated gene through the use of the proviral integration as a molecular “tag” at the site of insertion (Gaiano et al. 1996). Now that the zebrafish genome project is approaching completion, it is possible to isolate DNA fragments flanking the proviral integration on a large scale, sequence them and map the fragments to the proper location in the zebrafish genome, and then index the integration sites to cryo-preserved sperm samples. With this approach, a mutant line could be generated through *in vitro* fertilization of the frozen sperm sample containing an integration within the gene of interest (Wang et al. 2007). Here we have developed a new retroviral mutagenesis pipeline leveraging the power and cost efficiencies of a next-generation sequencing platform to isolate thousands of zebrafish gene mutations. We report the first 6,144 mutagenized and archived F_1 fish predicted to carry up to 3,776 mutations in zebrafish genes. The mutagenesis is ongoing and the mutagenized lines are being transferred to the Zebrafish International Resource Center (ZIRC) (Varga 2011) for open distribution to the research community. Large-scale mutagenesis of the zebrafish genome is the first step in defining the *in vivo* function of every gene in the zebrafish genome, and this retroviral mutagenesis resource complements other efforts in zebrafish to identify mutations using TILLING and gene trap technologies.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

RESULTS

We infected zebrafish founder fish with pseudotyped [M]MLV as described earlier (Wang et al. 2007). We generated >3,000 mosaic founder fish carrying multiple retroviral insertions. Each founder fish was then outcrossed with wild-type fish to obtain heterozygous F_1 fish**.** An average of six male F_1 fish per founder (ranging from 4 to 10 depending on the level of infection) were used to archive sperm samples and tail-biopsies was collected for insertion site identification. The outline of the approach used in this manuscript is shown in Figure 1. In our previous strategy, the most cost effective structure was to limit the sequences to 4 per fish (Wang et al., 2007). This maximized the number of *unique* sequences per fish, but capturing all the integrations was unlikely because of PCR amplification biases and limited sampling. Next generation sequencing platforms have the potential to overcome this limitation by making massive over-sampling of sequences inexpensive and therefore cost effective. We redesigned our mapping pipeline to take advantage of the depth of sequence afforded by next generation sequencing platforms.

Development of a high-throughput multiplexed mapping strategy

By adapting our mapping strategy to utilize Illumina HiSeq 2000 sequence data, we were able to significantly improve the efficiency of identifying the sequences flanking proviral integrations at a substantial reduction in cost. Each fish was receiving on average ≈390,000 sequences. This allowed us to use three frequently cutting restriction enzymes in parallel: (*MseI* T/TAA, *BfaI* C/TAG, and *Csp6I* G/TAC), rather than just the

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

single enzyme *Mse1*. Therefore if one restriction site was too close for mapping, the other sites provided additional chances for successful mapping.

Data generated using the new method captured a wider range of flanking sequences than did the original published single enzyme method. We adapted our linker-mediated PCR amplification protocol so that amplified genomic DNA adjacent to the retroviral integration sites could be directly sequenced on the HiSeq 2000 (Figure 2). To link a particular amplified fragment to the F_1 fish from which it originated, it was necessary to incorporate an index for the sequenced DNA fragments in the form of a 6-base "barcode" sequence adjacent to the ligation site of the linker (Figure 2). We synthesized 1,024 non-redundant linkers each containing a unique barcode, differing by at least two nucleotides between each bar code to avoid incorrect assignment by sequencing miscalls. This allowed us to multiplex hundreds of samples in one sequencing lane.

The Illumina HiSeq 2000 with "paired-end" reads provided up to 200 million paired sequencing reads per lane, with 8 lanes on a chip. The platform routinely generated sequencing reads of 101 bp from each end of a paired-end sequence, for a maximum of 202 bp of total sequence. Using the standard Illumina sequencing primers, we had to sequence through our LTR and linker primers, which were both 25bp long. We therefore obtained 76 bp of sequence from the viral LTR side, and 70 bp of genomic sequence plus the 6-bp barcode index from the linker end. We had a maximum of 146 bp of sequence used to map integrations, larger than the average genomic fragment length sequenced (based on restriction enzyme site availability, it is ≈65 bp) with 46% of the sequences having overlap between the paired-end reads.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Insertion site mapping strategy

We developed a new customized bioinformatics pipeline to map retroviral insertions in zebrafish genome. Each side of the raw sequence data was trimmed of LTR sequences or linker sequences. Barcodes for each fish were identified, indexed to the sequence but trimmed before alignment with the genome. We used two independent strategies to process the raw sequence data (contig construction or independent end mapping), followed by mapping retroviral integrations using the Bowtie algorithms (Langmead et al. 2009) (Figure 3, see supplemental figure 1 for details of mapping strategy). A "consensus" list of unique integration coordinates based on the two different mapping strategies proved to significantly eliminate mis-mapped integrations, allowing a better recovery rate of correctly mapped integrations after *in vitro* fertilization.

Zebrafish mutations generated from proviral integration

We utilized the new protocol to map retroviral insertions generated from founder injections. We processed F_1 genomic DNA samples from 6,144 male zebrafish (sixty-four 96-well plates) with matching cryo-preserved sperm using the overall approach described. Of the 6,144 F_1 fish, 15,223 unique integration sites (integrations with different genomic coordinates) were mapped using the consensus list from the two different mapping strategies. Among all insertions mapped to the genome, 52% of the integrations (7,896/15,223) were in genes annotated by Ensembl. Insertions in exons generate a truncation at the site of integration, and our mapping data shows 12% of the gene hits were in exons (963/7,896). 88% (6,933/7,896) of the gene "hits" were in introns. [M]MLV has a known bias to integrate near transcriptional start sites, and

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

further analysis of the introns hits show 40% of the integrations (2,813/6,933) that occurred in introns were in the first intron (Table 1 and Figure 4 A, B). This is consistent with the previous studies in mice (Mooslehner et al. 1990; Scherdin et al. 1990) and human tissue culture cells (Wu et al. 2003). Overall 60% (9,189/15,223) of all integrations landed either in genes or within 1 kb upstream or downstream of genes. The mapped integrations showed roughly equal distribution across all chromosomes (Figure 4 D). In order to determine if we could expect a broad distribution of identified mutations, we looked at the distribution pattern for predicted mutations. As expected, the vast majority of identified mutations have been hit only once (3,937), meaning we are still well below saturation for the technique. However, there are some clear hot spots for integration with 97 instances of genes hit independently more than five times (based on uniquely mapped coordinates) and five examples of genes hit more than 10 times (Figure 4 C). The overall integration distribution profile is consistent with our previous studies (Wu et al. 2003; Wang et al. 2007).

We have previously shown that integrations both in exons and in the first intron of genes are highly mutagenic (Wang et al. 2007). In 80% of genes that contained integrations in their first intron, mRNA levels were reduced to <10% of wild-type. Nearly one in five retroviral integrations is predicted to result in a disruption that reduces the gene expression level to 10% or less of the wild-type level (for the case of integrations in intron 1) or a truncation at the site of integration (for cases where the integration lands in an exon) (Wang et al. 2007). Extrapolating from our previous data, the number of potential null or severe hypomorph mutants in our data is up to 3,776. The number of unique genes for these mutations is 3,054.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Recovery of mapped insertions.

Using more conservative mapping methods we can recover alleles from frozen sperm stocks at up to a 79% success rate. We attempted to recover 197 alleles from our frozen samples. We confirmed 156 (79%) using PCR primers designed from sequences adjacent to the site of insertion. Another 4 were recovered based on primers designed to the raw sequencing data, but could not be confirmed to be in the correct genomic location suggesting that gene is not correctly mapped to the genome despite supposedly having a unique mapping position.

Phenotypic characterization of recovered alleles

Genetic screens focused on zebrafish morphology have typically been performed during the first five days of embryonic development. As expected, our retroviral insertional lines produced early embryonic phenotypes (Figure 5). To demonstrate the utility of this resource in producing biological information beyond embryonic development of zebrafish, we analyzed the nature of 41 mutations we recovered by *in vitro* fertilization through larval and adult lifespan (Supplementary Table 2). The F_2 fish (counting generations from the founder fish) were raised to adults then genotyped to identify mutant carriers. F_2 heterozygous mutant carriers were inbred and the embryos were examined for early developmental defects. 12 of the 41 showed obvious morphological defects in the developing embryo and all of them were genetically linked to the predicted mutagenic insertion. We tested embryos from each mutant line by semi-quantitative RT-PCR to determine if the mRNA for the predicted mutations were

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

affected and found that all of them showed significant reduction in mRNA transcript levels.

For all the mutations that did not have an early embryonic phenotype, we raised the F_3 generations and genotyped the fish once they reached sexual maturity. In most of the lines, we could detect the normal Mendelian ratios of 1:2:1, suggesting that the mutated genes did not have a significant impact on viability in a laboratory setting.

Figure 5 shows several phenotypes identified in the tested alleles. We show four examples of genes that showed early embryonic phenotypes: *wee1* mutants showed a very early cell death phenotype (Figure 5 A), *eif3s2* mutants showed defects in arterial patterning, *snapc1b* mutants showed liver and jaw defects, and *rpa* mutants showed a curled body axis and deficient brain and head structures. We had one allele (*Zgc:194470*) that had a juvenile morphological phenotype of overgrowth and two cases (*Slc7A5* and *Tg*) that had clear morphological phenotypes in adults. *Zgc:194470*$^{-/-}$ mutants were fully viable but had a larger body size by day 10-12 (Figure 5E), although eventually they became indistinguishable from their wild-type siblings. *Slc7A5*$^{-/-}$ adults were significantly smaller than their wild-type siblings and *Tg*$^{-/-}$ siblings had an enlarged red growth under their chins resembling a thyroid hypertrophy (Figure 5F,G) (Jao et al. 2008). These data demonstrate that this resource will generate not only mutations that would be readily identified as early developmental defects in forward screening, but can also identify genes that reduce viability, cause adult onset diseases, or alter adult morphology.

Distribution of mutants

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

After the integrations are mapped, the archived sperm samples are being deposited at the Zebrafish International Resource Center (ZIRC) for open distribution to scientific community.

Discussion

We have mapped 15,223 [M]MLV proviral integrations onto the zebrafish genome resulting in 3,776 predicted mutations in 3,054 genes (see Supplementary Table 1). By adapting the mapping pipeline from capillary sequencing to the HiSeq 2000, the number of sequences per F_1 fish increased from 4 sequences per fish to ≈390,000 sequences per fish. The extreme oversampling results in essentially saturating our ability to identify existing integrations (because of limited sampling and cost considerations, our previous approach only recovered approximately 20% of the existing integrations). Approximately 24% of all integrations are either in exons (6%) or in the first intron (18%) (Table 1). There were multiple examples of genes with more than one unique integration, suggesting that as number of identified integrations increases we will eventually reach a point where we will need to shift to an insertional DNA element with a different integration bias to reach genomic saturation.

We successfully recover nearly 80% of the retroviral integrations from the frozen sperm samples. The failures to recover mutations could reflect mistakes in the genome assembly, gaps in the genomic sequence where the integration would have mapped, or polymorphisms that result in a mis-alignment. In addition, some rate of human error cannot be discounted. We expect the recovery rate to improve as updated versions of the zebrafish genome are released.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

It is important to note the differences between our mapping approach compared to mutagenesis projects using [M]MLV previously undertaken in mouse embryonic stem cells. The mouse projects were very large-scale efforts, however, they relied on gene trap constructs and reporter expression (typically antibiotic resistance) to select for gene trap events. Thus the number of genes ultimately trapped by this approach cannot exceed the number of genes expressed at the time of retroviral infection and traps require both correct orientation and in-frame splicing events. Because of these limitations, typically gene traps do not end up trapping more that 50% of all genes in the genome. Because our approach does not rely on gene expression, but only on identifying the exact site of integration, it is likely that we will be able to mutagenize a significantly larger number of genes before the approach reaches saturation. Based on the vast majority of mutations we have identified so far having only one integration event (Figure 4 C), we believe we are still very far from saturation for this approach and we can continue to generate new mutations for several years. All frozen sperm samples are transferred to the Zebrafish International Resource Center. Each F_1 fish has four frozen samples. If particular mutations are requested multiple times, ZIRC will use one sample to raise multiple fish and re-freeze the samples. Making this a durable mutant resource.

Recently, transcription activator-like effector nucleases (TALENs) are being used as an alternative method for knock out genes in zebrafish (Huang et al. 2011; Bedell et al. 2012). While TALENs are an effective tool in for targeted mutagenesis, scaling this method to mutagenize thousands of genes would be very difficult, and no genome-wide resource utilizing TALENs is currently available.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

One of the key aspects of generating a defined collection of mutations in a broad spectrum of zebrafish genes is the utility of these alleles for generating models for human disease. Approximately 1/3 of the integrations we predict to be mutagenic were in exons (Table 1 and Figure 4 B), these are pure disruptions at the site of integration. The remaining alleles were in the first intron are typically severe hypomorphs based on measured mRNA levels. Extensive literature in *C. elegans*, *Drosophila*, and even mouse, have demonstrated that "weaker" alleles are often, if not in the majority of cases, superior to a null allele when analyzing gene effects in a multi-cellular organism. It is also worth mentioning that most human genetic diseases are hypomorphic mutations and pathology is identified well after the embryonic stage.

The indexing technique and mapping pipeline we developed for pooling samples into a single lane of the Illumina HiSeq2000 have important utility in a variety of research and clinical settings. The main advantages are: 1) a greater than ten-fold reduced cost compared to capillary sequencing with an ≈58% improvement in identifying integrations, 2) simple sample preparation that is amenable to scaling and automation. The technique does not require sonication of samples (Williams-Carrier et al. 2010) and has deeper sampling than 454 (Ciuffi et al. 2009). It is likely that the deeper sequencing is compensating for distortions caused by PCR amplification and other site cloning biases that may occur. The technique can be readily modified to map any DNA element being inserted into any sequenced genome. It has utility for mutagenesis using transposons such as *Tol2* or *Piggyback*, or in gene therapy experiments with any vectors that stably integrate into the genome.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

METHODS

Generation of Virus-infected Fish and Cryopreservation

Founder production, F_1 fish husbandry, and cryopreservation of sperm samples were performed as previously described (Wang et al. 2007). In brief, retroviral stocks were prepared according to Jao *et al.*(Jao and Burgess 2009), and synchronized embryos were obtained from wild-type *T/AB-5* fish. The concentrated viral stock was injected into blastula stage embryos (1000-2000 cell stage). Injected embryos were tested for the efficiency of proviral infection using qPCR-based assays to determine the copy number of the provirus (embryo assay values, EAV). High-quality founder fish with high EAVs were raised and F_1 fish were generated by outcrossing with wild-type fish. 5-10 F_1 male fish per founder were selected for cryopreservation and to map the retroviral integrations. The sperm from each F_1 fish line was collected and frozen, and the corresponding tail-cut was used to isolate genomic DNA for mapping.

Genomic DNA Preparation and Fragmentation

Genomic DNA was isolated from F_1 fish tail biopsies. In 96-well plates, each F_1 tail sample was lysed with 100µl of lysis buffer (10 mM Tris/HCl pH 8.0, 1mM EDTA, 50mM KCl, 2mM $MgCl_2$) 20 µg/µl proteinase K (Invitrogen, Inc. USA). After digestion at 55°C for 3 hrs, the DNA was precipitated with isopropanol, and washed with ethanol. The DNA pellet was dissolved in 50 µl of distilled water. We used approximately 500ng of genomic DNA for fragmentation using 3 pairs of restriction enzymes (*MseI/PstI, BfaI/BanII* (New England Biolabs, Inc. USA) *and Csp6I/ ECo24I* (Fermentas, Inc. USA))

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

in parallel. The restriction digestion was done at 37°C for 8 hours and heat inactivated at 80°C for 10 minutes.

Preparation of Barcoded Linkers

The barcode linkers followed a "splinkerette" design (Devon et al. 1995) with a 31 nucleotide long upper strand and 49 nucleotide long lower strand including a 6-nt barcode and a TA overhang. They were synthesized on 10 nm scale by IDT DNA Inc. The synthesized oligonucleotides were reconstituted in TE buffer to a 200µM concentration master stock. A 2 µM working concentration was prepared in STE buffer (TE with 50 mM NaCl). Barcode linkers were annealed at 70°C for 3 minutes, and 65°C for 10 minutes. A final concentration of 0.2 µM was used to ligate onto the restriction enzyme digested genomic fragments. The digested samples from each enzyme pair were pooled with pre-aliquoted barcoded linker in individual wells. The T_4 DNA ligase (New England Biolabs, Inc. USA) was added, and the reaction mix was incubated at 16°C for 6 hours.

Linker Mediated PCR

The linker mediated PCR was performed in two steps. In the first step, PCR was done with one primer specific to the 3'- LTR (5'-GACTTGTGGTCTCGCTGTTCCTTGG-3') and the other primer specific to linker sequences (5'-GTAATACGACTCACTATAGGGC-3') using the following conditions: 95°C for 2 minutes, 7 cycles of 95°C for 15 seconds, 72°C for 1 minute and then 32 cycles of 95°C for 15 seconds, 67°C for 1 minute and a final step for 4 minutes at 67°C. The PCR products were diluted to 1:50 in dH_2O, and a

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

second round of PCR was performed using LTR (5'-GAGTGATTGACTACCCGTCAGCGGGGGTCTTTCA-3') and Linker specific (5' -ACT ATAGGGCACGCGTGGTCGACTGCGCAT-3') nested primers to increase sensitivity and avoid non-specific amplification. The nested PCR products from each 96-well plate were pooled together and processed for Illumina library preparation as per manufacturer's instructions.

Illumina Library Preparation

Illumina libraries were prepared from 1 µg of pooled PCR products. Illumina paired-end adaptors were ligated onto LTR-gDNA-linker amplicons generated from the nested PCR reactions following Illumina's sample preparation guide. In brief, the PE adaptor oligo mix was incubated with PCR amplicons using T_4 DNA ligase at RT for 20 minutes. The ligation reaction was cleaned up using a QIAquick Min-elute column (QIAGEN) and eluted with EB buffer. The purified library was PCR-enriched with Phusion High-Fidelity polymerase in HF buffer. PCR was performed using primers (PE primer 1.0 and PE primer 2.0) supplied with the Illumina paired-end kit with the following conditions: 30 seconds at 98°C, 15 cycles of 98°C for 10 seconds, 65°C for 30 seconds, 72°C for 30 seconds, followed by a 5 minute at 72°C, and a final hold at 4°C. The PCR enriched library was purified using a QIAquick Min-elute column, and eluted in 20 µl of EB solution. An equi-molar concentration of different barcoded libraries were pooled together, and the final concentration was determined using quantitative PCR prior to loading onto the Illumina sequencer flowcell.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Retroviral Integration Mapping

Paired-end sequencing of multiplexed samples was performed on the Illumina GAIIx or HiSeq 2000 platforms. Sequence reads were extracted from the ELAND or BAM files generated by the sequencer. Non-zebrafish sequences were trimmed from each read. The six-base nucleotide "barcode" sequence was then identified and compared to a database of indexed sequence codes. The resulting trimmed sequences were mapped to the zebrafish genome (Ensembl Zv9 assembly build e65) using Bowtie. In order to increase confidence in the ability to recover a specific integration, integration sites with ≥30 redundant mappable sequence reads were selected as higher confidence. Two bioinformatics methods for processing and mapping sequences were used.

Mapping method one

Pre-processing and alignment of proviral insertion sites

We assembled a curated, single-ended library from the original paired end reads (Supplemental Figure 1). Each of the single sequences is assembled in the form 5'-Linker-barcode-flanking genomic sequence-LTR-3'. A brute force exact alignment algorithm (Castruita et al. 2011) was used to align the paired reads along their overlapping regions and to find the location of both the linker and LTR sequences. All sequences are then stored in the form: 5'-linker-barcode/ flanking genomic sequence/ LTR-3' for downstream analysis. Flanking sequences were extracted and aligned to the zebrafish genome assembly Zv9 using Bowtie (Langmead et al. 2009) with a tolerance of one mismatch. Only reads longer than 11 nt and with unambiguous alignments were used to pinpoint the insertion locus.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Some flanking sequences were sufficiently long that the paired reads did not overlap. In these cases, an oriented pseudo single end sequence was generated in the form 5'-Linker-barcode/ flanking genomic seq1/ N-flanking genomic seq2/ LTR-3'. The resulting flanking sequences were separately mapped to Zv9 using Bowtie (Langmead et al. 2009). Multiple hits were filtered to keep a maximum of 20 hits per read. In our model, the paired flanking sequences have a unique alignment if hits from both sequences are aligned to the same chromosomal region, same strand orientation, are at a distance of less than 1kb between hits, and there is only one hit-pair that meets the above requirements.

Identifying insertion sites

All unique hits (i.e. unambiguously mapped to a single location in the genome) from the pre-processing step were pooled, and integration coordinates were extracted from the Bowtie mapping output. The integration site was defined as the genomic coordinate immediately adjacent to the portion of the read to which the 3' LTR had been attached. The 3'-LTR end position was used to determine redundant sequences for each barcode or sample, and the longest fragment was used as representative to report and display the insertion locus.

For reporting and viewing the data, a bed-formatted file was produced, which give the chromosome, flanking sequence start, flanking sequence end, barcode, frequency (number of reads per integration site) and orientation of each integration and is available as a download from http://research.nhgri.nih.gov/ZInC/ (Varshney et al. 2013).

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Annotation of integration loci

The gene annotation file Ensembl ZV9 e65 was used to build a “One gene, One transcript” gene structure model (see Supplementary Figure 2) as the exonic union of all the annotated transcripts. Finally, bedtools was used to determine the overlap between the integration and the gene model, and integration sites were annotated as explained above.

Mapping Method Two

Trimming and orienting retroviral tag sequences, and barcode identification

Prior to mapping the reads, a custom script was run through to trim off the 3’ retroviral LTR and linker cassette (LC) sequences and identify the barcodes. Sequence reads were discarded if they did not contain either the 3’ LTR or the LC primer sequence in their 5’ end. The six nucleotides directly adjacent to the LC primer sequence represented the barcode. Sequences were trimmed of the 3’ LTR and/or LC sequence as well as the barcode, the barcode was noted for sample identification, and the trimmed sequence was used for mapping integrations.

Mapping retroviral tags

Bowtie (Langmead et al. 2009) was again used to map the trimmed retroviral sequence tags to the Zv9 zebrafish genome assembly, allowing for one mismatch. Since the site of retroviral integration and the sample barcode could occur on separate mate pairs, it was important to perform paired-end sequencing. However, sequence ends were mapped separately due to wide variation in both the trimmed sequence lengths and the

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

distances between the mate pairs. As above, the integration site was defined as the site of LTR insertion.

Pairing mapped sequence ends, collapsing redundant sequences, and identifying integration sites

After mapping, corresponding ends were paired and uniquely mapping read pairs were used to identify insertion sites. The following criteria were used for pairing and determining uniquely mapped insertion sites: both ends must map to the same chromosome within 1 kb of each other, and with the correct orientation (the 3' ends of the reads should point towards each other). Priority was given to the mapping that resulted in the smallest number of mismatches (supplemental figure 1B). The number of redundant sequences was recorded. Integration sites with ≥60 redundant sequence reads were used for downstream analyses.

Annotation of retroviral integration sites

The genomic position of retroviral integrations was compared to those of zebrafish gene models obtained from Ensembl Zv9 e65. A custom perl script(Hu et al. 2008) was run to identify those retroviral insertions that occurred within a gene or 1 kb upstream or downstream of a gene.

Data Access

All the sequence data from genomic DNA adjacent to the insertion sites used for mapping has been deposited in the public NCBI GSS database (BioSample ID: LIBGSS_038780). The detailed insertion data are also incorporated into ZFIN

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

(http://zfin.org), the zebrafish model organism database as transgenic insertions. To help researchers from other fields, we created a database, the Zebrafish Insertion Collection (ZInC) that can be searched using different search inputs such as human, mouse gene symbols or KEGG pathway terms (http://research.nhgri.nih.gov/ZInC)(Varshney et al. 2013). Insertion data can also be downloaded in bed file format to be used with the UCSC and Ensembl genome browser.

ACKNOWLEDGEMENTS

We would like to thank Colin Huck for superior animal care, and Darryl Leja for illustrations and graphics work, Robert Blakesley, Alice Young, and the National Intramural Sequencing Center (NISC) for providing sequence data. This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (SB, TW), R01 GM085357 and R01 HL109525 (AS), and R01 DK084349 (SL).

REFERENCES

Amsterdam A, Hopkins N. 1999. Retrovirus-mediated insertional mutagenesis in zebrafish. *Methods in cell biology* **60**: 87-98.

Bahary N, Davidson A, Ransom D, Shepard J, Stern H, Trede N, Zhou Y, Barut B, Zon LI. 2004. The Zon laboratory guide to positional cloning in zebrafish. *Methods in cell biology* **77**: 305-329.

Bedell VM, Wang Y, Campbell JM, Poshusta TL, Starker CG, Krug Ii RG, Tan W, Penheiter SG, Ma AC, Leung AY et al. 2012. In vivo genome editing using a high-efficiency TALEN system. *Nature*.

Castruita M, Casero D, Karpowicz SJ, Kropat J, Vieler A, Hsieh SI, Yan W, Cokus S, Loo JA, Benning C et al. 2011. Systems biology approach in Chlamydomonas reveals connections between copper nutrition and multiple metabolic steps. *The Plant cell* **23**(4): 1273-1292.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Ciuffi A, Ronen K, Brady T, Malani N, Wang G, Berry CC, Bushman FD. 2009. Methods for integration site distribution analyses in animal cell genomes. *Methods* **47**(4): 261-268.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.

Devon RS, Porteous DJ, Brookes AJ. 1995. Splinkerettes--improved vectorettes for greater efficiency in PCR walking. *Nucleic acids research* **23**(9): 1644-1645.

Friedel RH, Plump A, Lu X, Spilker K, Jolicoeur C, Wong K, Venkatesh TR, Yaron A, Hynes M, Chen B et al. 2005. Gene targeting using a promoterless gene trap vector ("targeted trapping") is an efficient method to mutate a large fraction of genes. *Proceedings of the National Academy of Sciences of the United States of America* **102**(37): 13188-13193.

Gaiano N, Amsterdam A, Kawakami K, Allende M, Becker T, Hopkins N. 1996. Insertional mutagenesis and rapid cloning of essential genes in zebrafish. *Nature* **383**(6603): 829-832.

Hu J, Renaud G, Gomes TJ, Ferris A, Hendrie PC, Donahue RE, Hughes SH, Wolfsberg TG, Russell DW, Dunbar CE. 2008. Reduced genotoxicity of avian sarcoma leukosis virus vectors in rhesus long-term repopulating cells compared to standard murine retrovirus vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**(9): 1617-1623.

Huang P, Xiao A, Zhou M, Zhu Z, Lin S, Zhang B. 2011. Heritable gene targeting in zebrafish using customized TALENs. *Nature biotechnology* **29**(8): 699-700.

Jao LE, Burgess SM. 2009. Production of pseudotyped retrovirus and the generation of proviral transgenic zebrafish. *Methods Mol Biol* **546**: 13-30.

Jao LE, Maddison L, Chen W, Burgess SM. 2008. Using retroviruses as a mutagenesis tool to explore the zebrafish genome. *Brief Funct Genomic Proteomic* **7**(6): 427-443.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Mooslehner K, Karls U, Harbers K. 1990. Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions. *J Virol* **64**(6): 3056-3058.

Mullins MC, Hammerschmidt M, Haffter P, Nusslein-Volhard C. 1994. Large-scale mutagenesis in the zebrafish: in search of genes controlling development in a vertebrate. *Curr Biol* **4**(3): 189-202.

Scherdin U, Rhodes K, Breindl M. 1990. Transcriptionally active genome regions are preferred targets for retrovirus integration. *J Virol* **64**(2): 907-912.

Solnica-Krezel L, Schier AF, Driever W. 1994. Efficient recovery of ENU-induced mutations from the zebrafish germline. *Genetics* **136**(4): 1401-1420.

Talbot WS, Schier AF. 1999. Positional cloning of mutated zebrafish genes. *Methods in cell biology* **60**: 259-286.

Varga ZM. 2011. Aquaculture and husbandry at the zebrafish international resource center. *Methods in cell biology* **104**: 453-478.

Varshney GK, Huang H, Zhang S, Lu J, Gildea DE, Yang Z, Wolfsberg TG, Lin S, Burgess SM. 2013. The Zebrafish Insertion Collection (ZInC): a web based,

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

searchable collection of zebrafish mutations generated by DNA insertion. *Nucleic acids research* **41**(D1): D861-864.

Wang D, Jao LE, Zheng N, Dolan K, Ivey J, Zonies S, Wu X, Wu K, Yang H, Meng Q et al. 2007. Efficient genome-wide mutagenesis of zebrafish genes by retroviral insertions. *Proceedings of the National Academy of Sciences of the United States of America* **104**(30): 12428-12433.

Waterston RH Lindblad-Toh K Birney E Rogers J Abril JF Agarwal P Agarwala R Ainscough R Alexandersson M An P et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562.

Williams-Carrier R, Stiffler N, Belcher S, Kroeger T, Stern DB, Monde RA, Coalter R, Barkan A. 2010. Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. *The Plant journal : for cell and molecular biology* **63**(1): 167-177.

Wu X, Li Y, Crise B, Burgess SM. 2003. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**(5626): 1749-1751.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Figure legends

Figure 1. Overview of the retroviral mutagenesis pipeline

The pseudotyped Murine Leukemia Virus (MLV) is injected into 1000-2000 cell stage blastula embryos. The infection rate is determined by quantitative PCR (qPCR) and founder fish with high infection rates are raised to adults. The founders are crossed to wild-type (T/AB) fish and F_1 male fish are used for sperm cryopreservation and fin biopsies. Integrations are amplified and mapped from gDNA isolated from the fin biopsies. Mapped integrations are assigned to the corresponding sperm samples, and desired mutations are recovered by *in vitro* fertilization.

Figure 2. Overview of high-throughput strategy to identify retroviral integrations using a next-generation sequencing platform

Genomic DNAs corresponding to individual F_1 fish were digested with three sets of restriction enzymes in parallel. After heat-inactivation of the restriction enzymes, the digested samples were then pooled together and ligated with DNA linkers, each containing a unique 6-bp barcode that indexes the F_1 fish. The linker ligated DNA fragments were amplified by linker-mediated PCR using linker and viral LTR specific primers to amplify the adjacent genomic DNA sequences. The LTR/gDNA/linker amplicons are subsequently ligated to Illumina paired end adapters and sequenced using the Illumina sequencing platform.

Figure 3. Strategies for mapping retroviral integrations

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Paired-end sequencing was performed to capture the site of the retroviral integration (designated by LTR – retroviral 3' long terminal repeat) and the linker cassette (LC) that contains the "barcode" identifier for the specific sample. Two strategies were used to map the integrations as this proved to be less error-prone than either strategy alone. In Strategy A, pairwise alignment of paired-end reads was performed to create contigs, and the resulting contigs were mapped to the zebrafish genome. Only contigs that mapped unambiguously were considered for identifying integrations. In Strategy B, each read from corresponding paired-ends was mapped independently and co-localization in the correct orientation (pointing at each other) was used as the criterion for correct mapping. Integrations that mapped to the same genomic coordinates by both strategies were used for identification of integration events.

Figure 4. Summary of proviral integrations from 6,144 F_1 fish

(A) Distribution of the 6,933 retroviral integrations in introns; 40% of integrations (2,813/6,933) integrations are in the first intron. (B) Distribution of 963 integrations in exons. (C) Number of hits per gene **based on integrations with unique genomic coordinates**, 72% of genes have only one integration. (D) Distribution of 15,223 integrations across all chromosomes.

Figure 5. Representative embryonic, larval and adult phenotypes from selected retroviral insertional alleles.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

(**A**) Insertion in the *wee1* gene led to an early phenotype of cellular necrosis starting at the gastrulation stage. Images here show a wilt type and mutant embryo at the 12 somite stage.

(**B**) Insertion in the *eif3s2* gene led to a vascular defect in homozygous mutants. The upper panel shows bright field images and the lower panel shows the lack of intersegmental vessels labeled by the *flk-gfp* transgenic marker in the *eif3s3*$^{-/-}$ background at 1dpf. (**C**) An insertion in the *snapc1b* gene causes embryonic phenotypes including jaw defects and a small liver visible at 5dpf. Arrows point to the reduced jaw structures in the mutant, dashed lines demarcate the liver. (**D**) Homozygous *rpa1* mutants at 2dpf have small and necrotic heads, small eyes, and tails curling dorsally. These homozygous phenotypes are weaker but observable at 1 dpf. All homozygotes die at approximately 5 dpf. **(E)** Insertion in a novel gene (*zgc:194470*) led to the larval phenotype of a larger body at day12 of development. **The mutant is homozygous viable and the body sizes become same as that of wild type when reaching the adult stage. 100% of the homozygous mutants show the larger larval phenotype (N=200).** **(F)** *Slc7a5*$^{-/-}$ fish showed no observable embryonic defects but they are 40% smaller than their wild-type or heterozygous siblings at 4 months of age. *Slc7a5* is a small subunit of the L-type amino acid transporter 1. **(G)** 6-month old adult *tg*$^{-/-}$ (thyroglobulin) fish showed red swelling under the chins (black arrows), a phenotype reminiscent of human thyroid goiters. *Tg*$^{-/-}$ fish are fertile and showed no observable embryonic defects.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Supplementary Figure 1. Workflow of two mapping methods.

(A) The assembly of a single-end curated library from an insertional paired-end library. Given the original paired end library, we seek to build a single end library of reads with a fixed Linker-LTR orientation and barcode identification. A brute force exact alignment method is employed to compute the best overlap between the first and second end of each pair, and for the location of linker and LTR sequences. In short, each sequence is transformed into a matrix representation so that their convolution provides the number of common bases for any possible overlap between them. Those assembled single-end sequences that with a length of less than 11nt were discard.

(B) Each sequence read from paired-end sequencing was mapped independently. The genomic position and the alignment quality for each corresponding paired reads were then used to determine unambiguously mapping events. For paired reads where both ends had alignment hits, if those reads mapped on the same chromosome, within 1 kb, and in the correct orientation at a single locus, the read was considered uniquely mapping. If read pairs failed to map unambiguously according to these criteria, single-end reads that mapped uniquely were also retained for identifying potential integrations.

Supplementary Figure 2. Annotation of integration events.

To determine the location of integration within a gene, a composite transcript was created from the annotated zebrafish transcripts from ENSEMBL (e65). All transcripts annotated in a given locus are consolidated into a single model transcript. The final representative gene model A' is the exonic union of all the annotated transcripts for gene A and is used for mutagenesis prediction.

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

Table 1. Distribution profile of 15,223 retroviral integrations in 6,144 F_1 fish

	Number of integrations	Percentage of Total Integration
Exon hits	963	6.3%
Total Intron hits	6,933	45.5%
First Intron hits	2,813	18.4%
500 bp upstream/downstream hits	780	5.1%
1000 bp upstream/downstream hits	1,293	8.4%

Fig 1

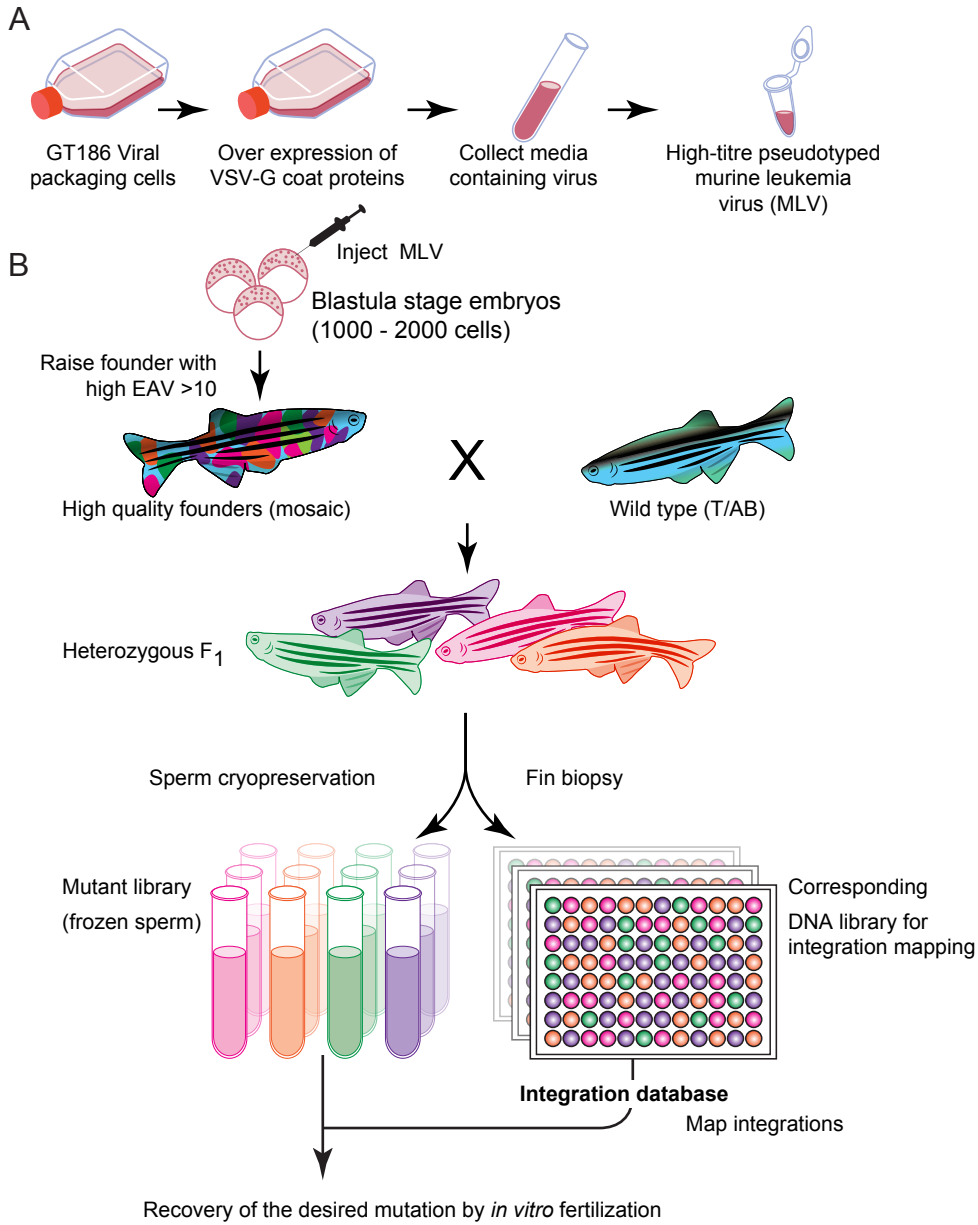

A
GT186 Viral packaging cells
Over expression of VSV-G coat proteins
Collect media containing virus
High-titre pseudotyped murine leukemia virus (MLV)
B
Inject MLV
Blastula stage embryos (1000 - 2000 cells)
Raise founder with high EAV >10
High quality founders (mosaic)
X
Wild type (T/AB)
Heterozygous F1
Sperm cryopreservation
Fin biopsy
Mutant library (frozen sperm)
Corresponding DNA library for integration mapping
Integration database
Map integrations
Recovery of the desired mutation by *in vitro* fertilization

Fig 2

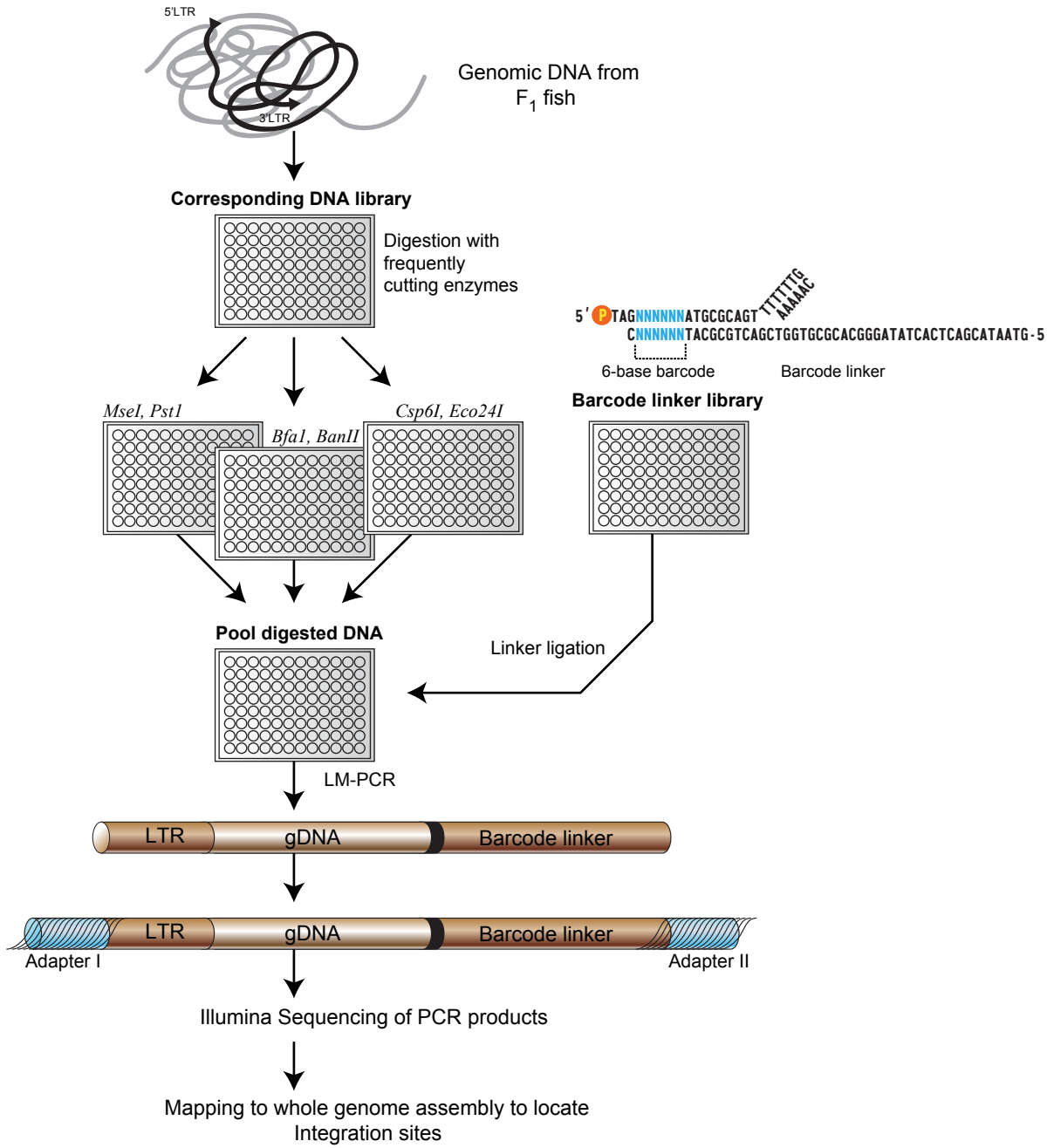
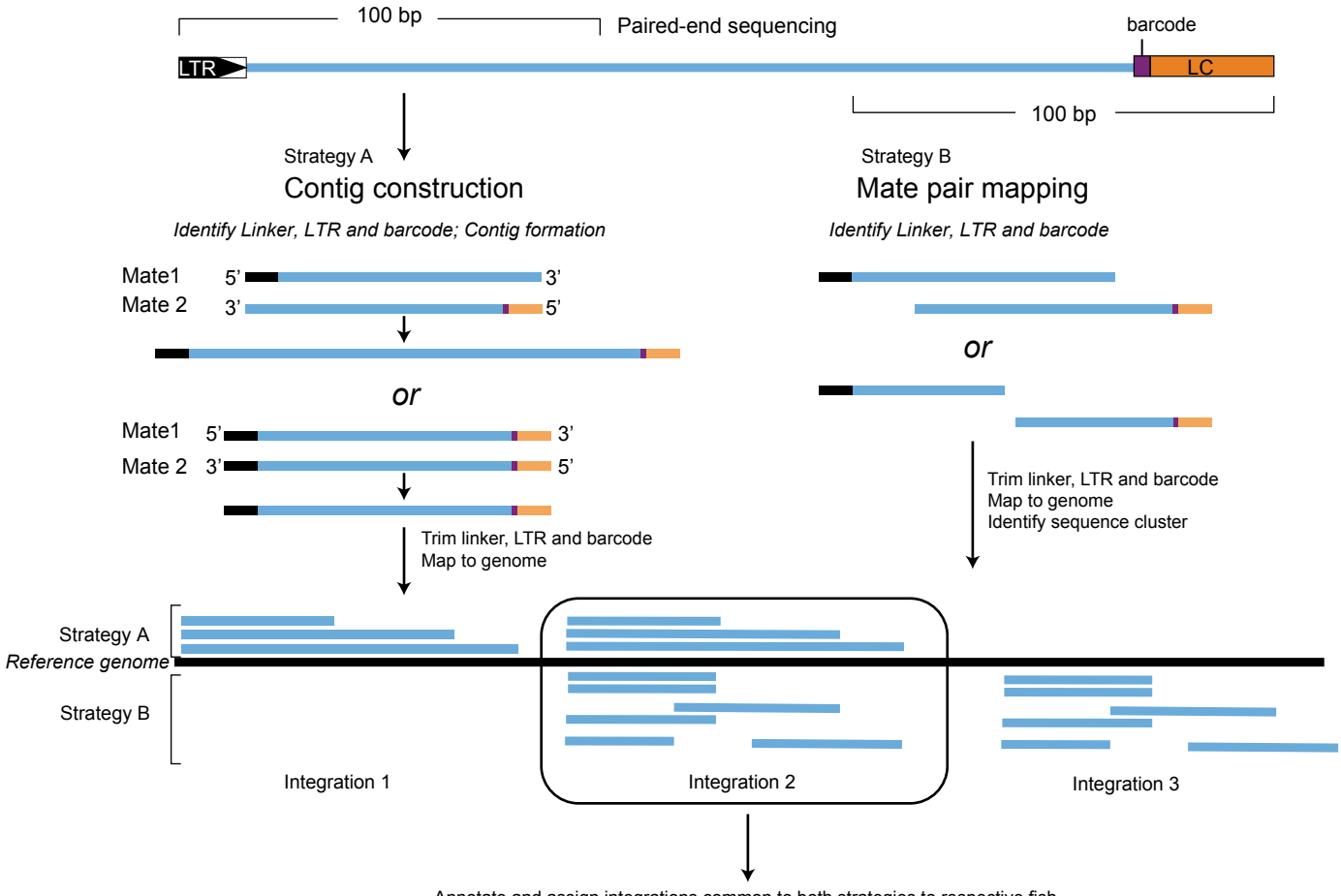
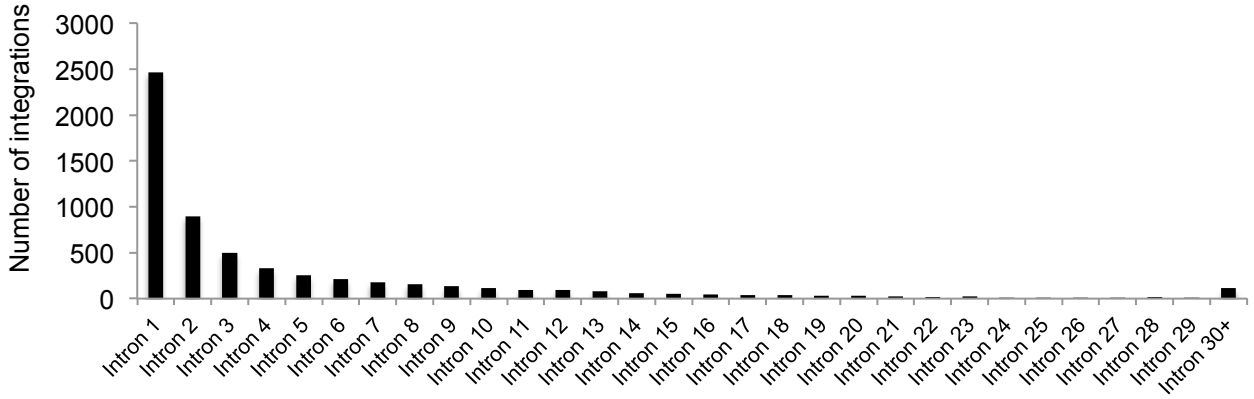

5'LTR
3'LTR
Genomic DNA from
F1 fish
Corresponding DNA library
Digestion with
frequently
cutting enzymes
5′ P TAGNNNNNNATGCGCAGT TTTTTTG AAAAAC
CNNNNNNTACGCGTCAGCTGGTGCGCACGGGATATCACTCAGCATAATG-5′
6-base barcode
Barcode linker
Barcode linker library
MseI, Pst1
Bfa1, BanII
Csp6I, Eco24I
Pool digested DNA
Linker ligation
LM-PCR
LTR
gDNA
Barcode linker
LTR
gDNA
Barcode linker
Adapter I
Adapter II
Illumina Sequencing of PCR products
Mapping to whole genome assembly to locate
Integration sites

Fig 3

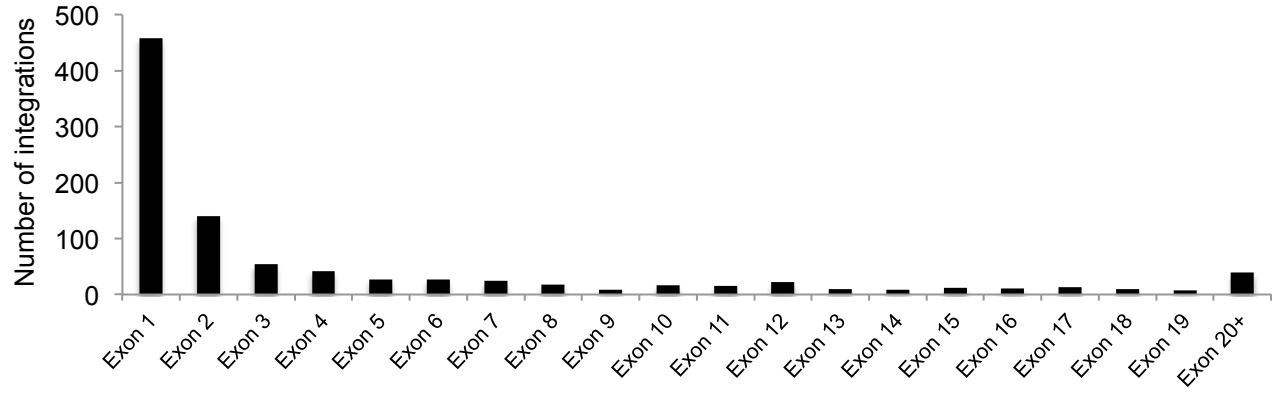

100 bp
Paired-end sequencing
barcode
LTR
LC
100 bp
Strategy A
Contig construction
Identify Linker, LTR and barcode; Contig formation
Mate1 5' 3'
Mate 2 3' 5'
or
Mate1 5' 3'
Mate 2 3' 5'
Trim linker, LTR and barcode
Map to genome
Strategy B
Mate pair mapping
Identify Linker, LTR and barcode
or
Trim linker, LTR and barcode
Map to genome
Identify sequence cluster
Strategy A
Reference genome
Strategy B
Integration 1
Integration 2
Integration 3
Annotate and assign integrations common to both strategies to respective fish

Downloaded from genome.cshlp.org on February 8, 2013 - Published by Cold Spring Harbor Laboratory Press

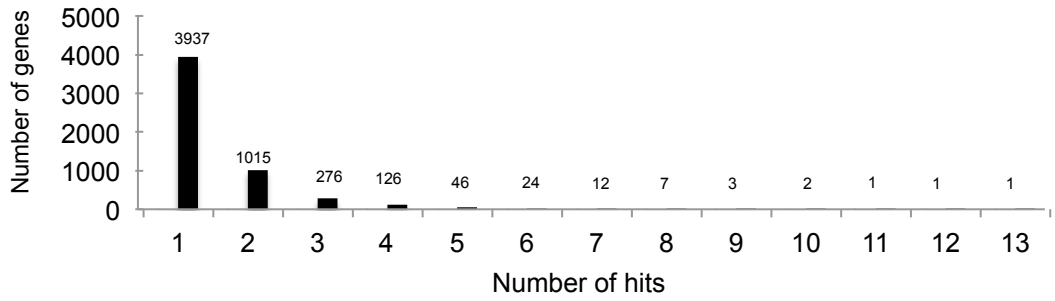
A



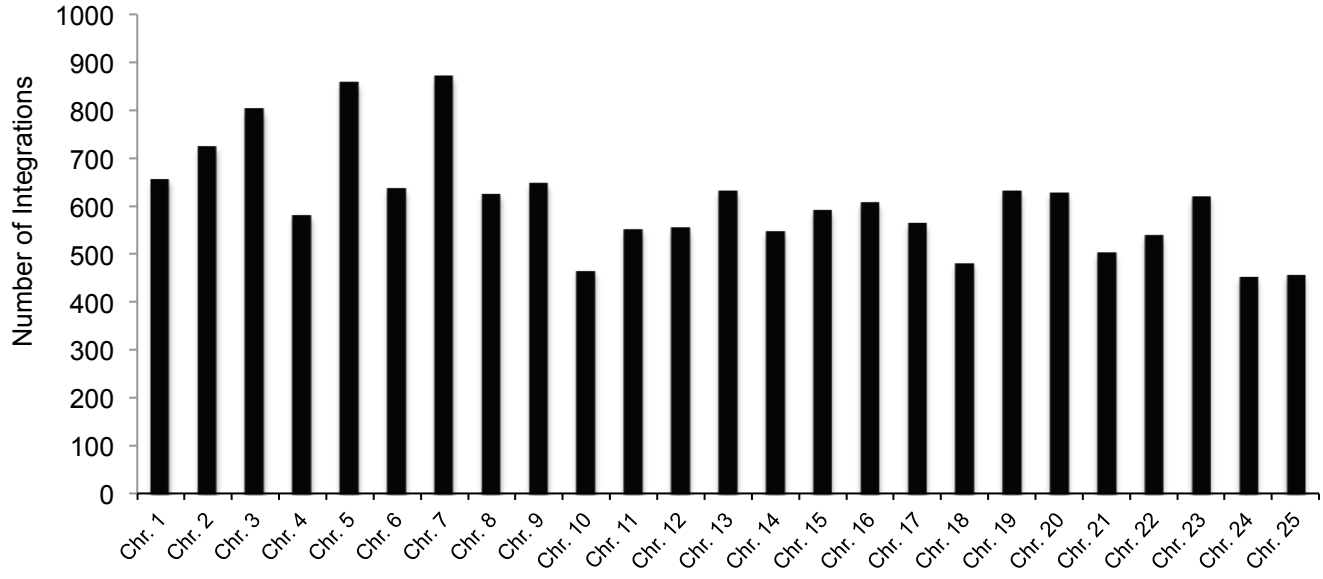
B

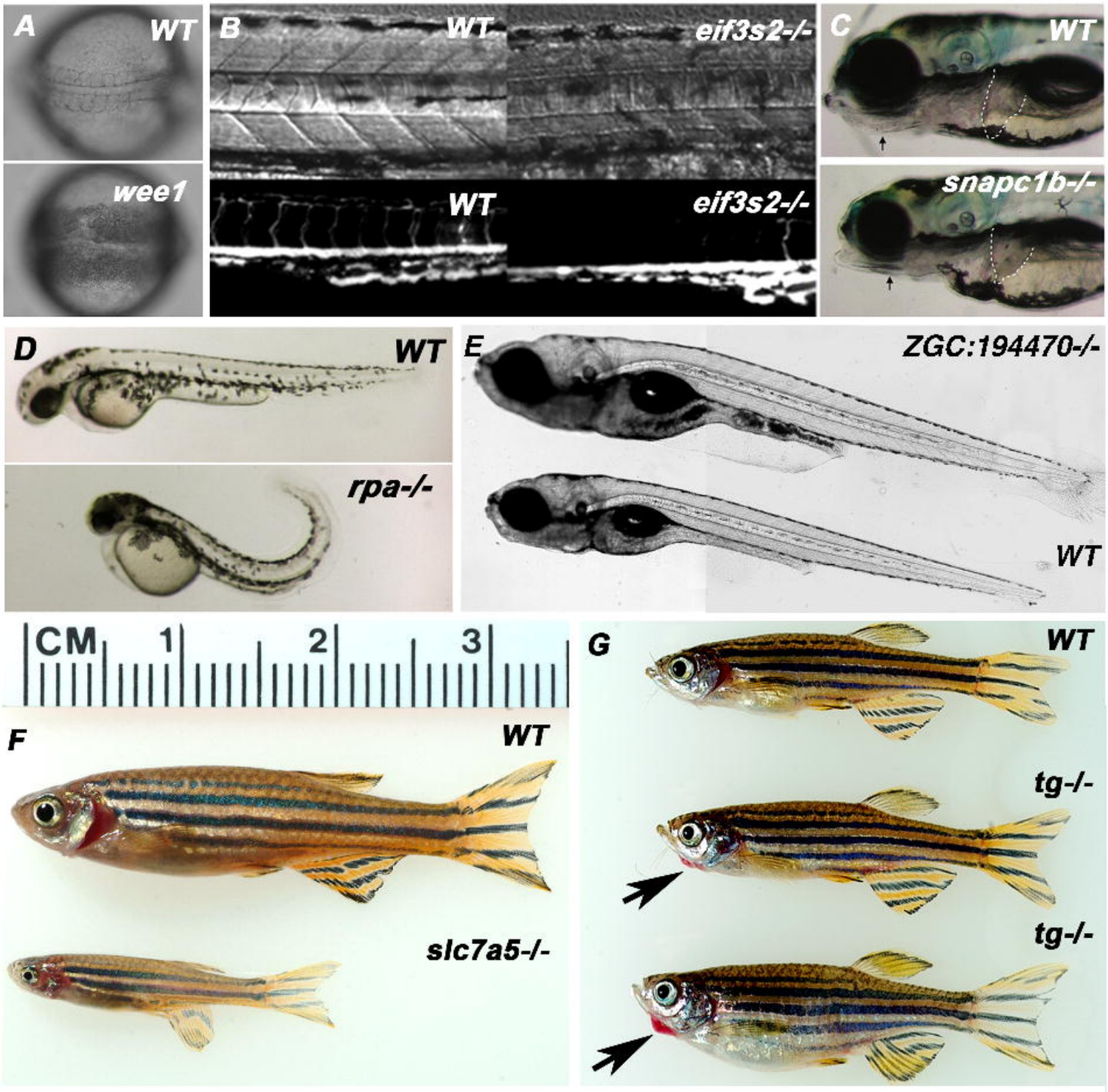


C



D




A
WT
wee1
B
WT
eif3s2-/-
WT
eif3s2-/-
C
WT
snapc1b-/-
D
WT
rpa-/-
E
ZGC:194470-/-
WT
CM
1
2
3
F
WT
slc7a5-/-
G
WT
tg-/-
tg-/-