Article:

Kettleborough, R.N.W., Busch-Nentwich, E.M., Harvey, S.A. et al. (17 more authors) (2013) A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature, 496. pp. 494-497. ISSN 0028-0836

https://doi.org/10.1038/nature11992

Published in final edited form as:
Nature. 2013 April 25; 496(7446): 494–497. doi:10.1038/nature11992.

A systematic genome-wide analysis of zebrafish protein-coding gene function

Ross N. W. Kettleborough[1,*], **Elisabeth M. Busch-Nentwich**[1,*], **Steven A. Harvey**[1,*], **Christopher M. Dooley**[1], **Ewart de Bruijn**[3], **Freek van Eeden**[2], **Ian Sealy**[1], **Richard J. White**[1], **Colin Herd**[1], **Isaac J. Nijman**[3], **Fruzsina Fényes**[1], **Selina Mehroke**[1], **Catherine Scahill**[1], **Richard Gibbons**[1], **Neha Wali**[1], **Samantha Carruthers**[1], **Amanda Hall**[1], **Jennifer Yen**[1], **Edwin Cuppen**[3], and **Derek L. Stemple**[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK [2]MRC-CDBG/Department of Biomedical Science, The University of Sheffield, Western Bank, Sheffield, S10 2TN, UK [3]Hubrecht Institute, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT, Utrecht, The Netherlands

Abstract

Since the publication of the human reference genome, the identities of specific genes associated with human diseases are being discovered at an enormous rate. A central problem is that the biological activity of these genes is often unclear. Detailed investigations in vertebrate model organisms, typically mice, have been essential for understanding the activities of many orthologues of these disease-associated genes. Although gene-targeting approaches[1-3] and phenotype analysis have led to a detailed understanding of nearly 6,000 protein-coding genes[3,4], this number falls significantly short of all >22,000 mouse protein-coding genes[5]. Similarly, in zebrafish genetics, one-by-one gene studies using positional cloning[6], insertional mutagenesis[7-9], antisense morpholino oligonucleotides[10], targeted re-sequencing[11-13] and zinc finger and TAL endonucleases[14-17] have made significant contributions to our understanding of the biological activity of vertebrate genes, but the number of genes studied again falls well short of the >26,000 zebrafish protein-coding genes[18]. Importantly, for both mice and zebrafish, none of these strategies is particularly suited to the rapid generation of knockouts in thousands of genes and the assessment of their biological activity. Enabled by a well-annotated zebrafish reference genome sequence[18,19], high-throughput sequencing and efficient chemical mutagenesis, we describe an active project that aims to identify and phenotype disruptive mutations in every zebrafish protein-coding gene. Thus far we have identified potentially disruptive mutations in more than 38% of all known protein coding genes. We have developed a multi-allelic phenotyping scheme to efficiently assess the effects of each allele during embryogenesis and have analysed the phenotypic

Correspondence should be addressed to D.S. (ds4@sanger.ac.uk) or E.C. (e.cuppen@hubrecht.eu)..
[*]These authors contributed equally to this work.

Author Contributions R.K., E.B.N., S.H. initiated, organised and executed the work described here in equal contribution. R.K., E.B.N., S.H. and D.S. designed the experiments. R.K., E.B.N. and S.H. wrote the manuscript with assistance from I.S., R.W., C.D. and D.S. Mutagenesis was performed by R.K., F.V.E. and E.D.B. The mutation analysis pipeline was developed by I.S. and I.J.N and maintained by R.W., C.H and F.F. implemented and improved genotyping procedures, F.F. and E.B.N. developed cryopreservation procedure, S.H., S.M., C.S., C.D., N.W. performed the phenotyping, J.Y. designed and tested the first Agilent SureSelect exome set, R.G. helped maintain and distribute alleles. S.C. and A.H. provided assistance for cryopreservation and genotyping. E.C. and D.S. collaborated in the initiation, design and process development of the project. All authors read the manuscript and provided comments.

Author Information Detailed information on alleles and their availability can be found on the web portal www.sanger.ac.uk/Projects/D_rerio/zmp/. All sequencing data is deposited in www.ebi.ac.uk/ena/data/view/ERP000426. The authors declare no competing financial interests.

consequences of over 1000 alleles. All mutant alleles and data are available to the community and our phenotyping scheme is adaptable to phenotypic analysis beyond embryogenesis.

Over the past nine years we have been establishing methods for the systematic identification of disruptive mutations in zebrafish. Given the lack of complete annotation at early stages of the project we originally employed a reverse genetic approach known as TILLING to identify mutations in specific genes by sequencing PCR-amplified exons from thousands of *N*-ethyl-*N*-nitrosourea (ENU) mutagenised individuals[11-13,20]. With the advent of high-throughput sequencing methods we were able to significantly increase the throughput of this approach but never reached genome-wide coverage of exons[11].

With the release of the Zv8 and Zv9 assemblies of the zebrafish genome and their protein annotations, we were able to design reagents to extract the annotated exons from zebrafish genomic DNA, enriching for ~60 Mbp of exome sequence and covering all 26,206 protein-coding genes[18,19] (Fig. 1). By sequencing the exon-enriched DNA from mutagenised F1 individuals we identify ENU induced mutations using a modified version of the 1000 Genomes project variant calling pipeline[21,22]. We increased throughput by pre-capture pooling up to eight bar-coded F1 genomic libraries and combining the exon-enriched DNA into a single high-throughput sequencing sample while retaining adequate sequencing coverage and depth to identify heterozygous induced mutations (Supplementary Table 1). For each F1 we now predict protein-coding consequences of the induced mutations and for each nonsense and essential splice-site mutation we generate a single nucleotide polymorphism (SNP) genotyping assay[11] to facilitate the identification of each mutation in subsequent generations. We are able to confirm 95% of candidate mutations in subsequent generations.

We have analysed the exome sequences of 1673 mutagenised F1 individuals and identified 12,002 induced nonsense and 5337 induced essential splice mutations in 10,043 genes. For 4105 genes we have identified two or more disruptive alleles (Supplementary Table 2). With this set of data we can make predictions about the number of F1 individuals we will need to analyse to obtain disruptive mutations in each protein-coding gene. The detection of nonsense and essential splice mutations in new genes gradually decreased over the 1673 sequenced individuals (Fig. 2a), such that the ratio of mutations in new genes to alleles was 1 for the first ten sequenced exomes, but 0.37 for the last ten exomes. On average, each individual contained 125 nonsense and 168 essential splice mutations, which were common to the strains used, and among induced mutations were 7 nonsense, 3 essential splice and 90 non-synonymous mutations (Supplementary Table 3). As the number of induced non-synonymous mutations within each individual was ~10X the sum of nonsense and essential splice mutations (Fig. 2b), the rate of detecting non-synonymous mutations in new genes decreased more rapidly over the 1673 exomes, with the ratio of non-synonymous mutations in new genes to alleles being 0.024 for the last ten exomes. Sequencing 808 individuals resulted in the identification of 85,338 non-synonymous alleles, corresponding to mutations in 75% of all known protein-coding genes. We predict that we will identify at least one disruptive mutation in 75% of protein-coding genes by sequencing ~8000 F1 individuals.

To yield the most value from this resource it is important to identify any phenotypic consequence of homozygous mutations. Thus we have established a high-throughput, systematic phenotypic analysis of alleles to assess the developmental consequences of any given mutation. We have focussed our efforts on induced nonsense and essential splice mutations. In a two-step, multi-allelic, phenotyping approach, we first identify those mutations that do not cause a phenotype in the F3 embryos at 5 days post fertilisation (dpf) (Fig. 3) resulting from crosses of up to 12 pairs of F2s (Fig. 3b). We genotype phenotypically normal 5 dpf F3 embryos for all mutations identified as heterozygous in both

F2 parents (Fig. 3c). Homozygous mutations present in the expected Mendelian ratios among F3 embryos are documented as not causing a phenotype at 5 dpf. Homozygous mutations present in less than 25% of phenotypically normal embryos are suspected to cause a phenotype (Fig. 3c).

In the second step of analysis, we test for correlations between the predicted disruptive mutations and morphological phenotypes (Fig. 3d-e). We re-cross the F2 adults that are heterozygous for the suspected causal mutation and examine the F3 embryos for all morphological and behavioural phenotypes during the first 5 dpf. All phenotypes are genotyped for the given mutation and if over 90% of embryos, for a given phenotype, are homozygous for the specific mutation it is documented as likely to be causal (Fig. 3d-e), with 10% tolerance for pipetting and genotyping errors. If less than 90% of embryos with one phenotype are homozygous for the mutation of interest, it is documented as being linked to a phenotype rather than causative.

We have performed a phenotypic analysis of 1216 nonsense and essential splice mutations. Of these, 48 mutations caused a phenotype within the first 5 dpf and 77 alleles were linked to a phenotype. Among the predicted disruptive mutations, 1065 were deemed to have no phenotype at 5 dpf and 26 are under further investigation. For all phenotype-genotype correlations we annotated each of the phenotypic traits using developmental stage, anatomical entity and phenotypic quality terms to enable phenotype data mining.

This mutation discovery and multi-allelic phenotyping pipeline systematically annotates zebrafish gene function. Importantly, the described genotype and phenotype correlations do not constitute proof of causality for the individual allele. Detailed aetiology of a phenotype-genotype correlation can only be proven by more exhaustive investigation, such as a complementation test. By combining two distinct potentially disruptive mutations in the same gene using a compound cross of two independent heterozygous carriers and a genotype analysis of the expected single-phenotype F3 embryos we can rule out linked mutations independently associated with each allele. Where a phenotype and allele have been associated and additional alleles in the same gene are available we perform complementation crosses to prove causality (Fig. 4).

We find about 6% (74/1216) of alleles are phenotypic, which is low by comparison to measurements of mouse embryonic lethality[23]. There are several possible explanations. Firstly, the alleles generated and analysed in this project are random point mutations across the length of each gene. Therefore a proportion of alleles do not or only partially disrupt protein function. However, the position of a mutation is not necessarily a good predictor for the severity of disruption as is demonstrated by the alleles described in Figure 4 d, h, k. Secondly, our phenotyping assays include only those morphological and behavioural changes detectable during the first 5 dpf in live embryos. Subtle phenotypes that require further intervention, such as immunohistochemistry, are not currently assayed. Finally, the teleost-specific genome duplication might cause paralogue redundancy. While this is possible there are few examples of paralogues being completely redundant. In contrast, there are numerous paralogue pairs where gene expression domains and functions have split between paralogues[24,25]. For example, mutants in *ttna* (Fig. 4 i-k) show a phenotype distinct from the published *ttnb* mutant *runzel* [26].

It is unlikely that a comprehensive functional understanding of all human protein-coding genes will become available in the near future. Therefore, by providing a systematic analysis of zebrafish gene function, with phenotypes annotated in searchable, ontology-based datasets, the reagents described here will advance our knowledge of the biological basis for human disease. So far we have identified mutations in the orthologues of 3188 of the 5494 genes currently associated with human disease in GWAS studies (www.genome.gov/gwastudies) and have at least one allele in 2505 of the 4204 genes associated with a human phenotype in OMIM (www.omim.org).

Our analysis will provide a rich resource for developmental biologists and clinicians to facilitate the identification of candidate genes for idiopathic inherited diseases or pathogen susceptibility. Furthermore, the alleles and data generated by this project are available to the scientific community through our website (www.sanger.ac.uk/Projects/D_rerio/zmp) and alleles will be available from two international stock centres, the Zebrafish International Resource Center (ZIRC) (www.zebrafish.org/zirc) and the European Zebrafish Resource Center (EZRC) (www.itg.kit.edu/ezrc). Information on how to use these facilities can be found in the supplementary information. We believe the work described here will significantly enhance the use of zebrafish as a model organism to study development and human disease.

Methods

Exome sequencing and SNP calling

Adult male zebrafish were mutagenised using ENU according to improved mutagenesis protocols[11]. G0 mutagenised individuals were outcrossed to create large F1 mutagenised libraries. DNA was isolated from F1 individuals by incubating fin biopsies in 400 μl of 100 μg/ml proteinase K for 10 h at 55°C, followed by 15 min at 85°C to heat inactivate the proteinase K. DNA was precipitated by adding 400 μl of isopropanol and centrifuging for 30 min at 4000 rpm and 4°C. DNA pellets were washed twice with 400 μl of 70% ethanol followed by centrifugation at 4000 rpm for 5 min, and resuspended in ddH_20. DNA from each individual (1-2 μg) was sheared and used to construct 150-200 bp insert Illumina libraries according to the manufacturers standard protocols.

For the rapid identification of ENU-induced mutations in individual zebrafish covering all annotated zebrafish protein-coding genes, we developed a whole exome enrichment reagent using Agilent SureSelect. Oligonucleotides (120mers) were designed at 2X tiling across the predicted exon coordinates and then manufactured as biotinylated RNA baits and blended into one tube ready for enrichment. Following completion of a pilot experiment to evaluate the technology, we developed an exome design using the Ensembl 61 (Zv9) gene set, which included a total of 60 Mbp of coding sequence and 26,206 genes[18,19].

For each F1 genomic Illumina library, 500 ng of DNA was hybridised for 24 h to biotinylated whole exome RNA baits. Hybridised fragments were enriched using streptavidin-coated beads, RNA was digested, and remaining libraries of fragments were amplified for 10 cycles using standard Illumina primers with or without indexing barcodes. The resulting amplified libraries were run on Illumina GAII or HiSeq2000 machines using GAII, HiSeq2000v2 or HiSeq2000v3 chemistries to perform 54 bp paired-end sequencing.

Initially, each enriched sample was sequenced on an individual lane of an Illumina GAII machine using 54 bp paired end runs, achieving a mean of 64 million reads per sample (Supplementary Table 1). Of those reads, 55% mapped to the exome target sequence, with 90% of the target being covered at 4X and 64% covered at 20X. For the identification of single nucleotide polymorphism (SNP) variants, we required at least 4X coverage, with 20X coverage providing the number of reads required for reliable mutation detection[29]. We subsequently moved to the HiSeq platform, incorporated barcoding into the library making and performed pre-capture pooling of libraries. These improvements allowed us to sequence 8 exomes on an individual lane, consequently increasing the throughput and lowering costs (Supplementary Table 1). These results demonstrated that we could efficiently enrich and sequence the zebrafish exome at the coverage required for reliable mutation detection in a cost-effective manner.

We identified ENU-induced mutations within the exome sequences using a modified version of the 1000 Genomes Project variant calling pipeline[21]. Paired reads were aligned to the Zv9 reference assembly using BWA and SNPs were called by SAMtools mpileup, QCALL and the GATK Unified Genotyper. SNPs not called by all three callers were removed from the analysis, along with any SNP that did not pass a caller's standard filters. Additionally, SNPs were removed where the total read depth was less than the number of samples and where the genotype quality was lower than 100 for GATK and lower than 50 for QCALL and SAMtools mpileup. Finally, SNPs within 10 bp of an indel (called by both SAMtools mpileup and Dindel) were removed. Variation consequences were assigned by the Ensembl Variant Effect Predictor[30] using the Ensembl 61 gene set. A SNP was defined as an induced mutation if present in one to three individuals, as this allowed for founder effects that may have arisen from the mutagenesis. If the SNP was present in more than three individuals it was considered to be common to the strain used. Heterozygous calls found in one to three samples only were deemed to be induced mutations and those with a nonsense consequence (Sequence Ontology accession SO:0001587) or essential splice consequence (SO accessions: SO:0001574 and SO:0001575) were used to design KASP assays.

In a pilot run, a sample of six individual exomes, we confirmed all induced mutations that were called by at least two of the SNP callers by both KASP genotyping and capillary sequencing to calculate the false positive rate of our analysis (data not shown). These genotyping results were used to set filtering parameters within our SNP calling pipeline, such that 85.7% of SNPs that could not be confirmed by KASP genotyping or capillary sequencing were removed. Information on common SNPs and insertions/deletions were also collected in order to avoid them when designing genotyping assays.

To estimate the mutation saturation we used missense mutations, which are present at a tenfold higher rate. As each mutagenised library displayed different rates of mutagenesis the order that exomes were sequenced was randomised in the analysis.

We reasoned that the induced nonsense and essential splice mutations were more likely to result in putative loss of function alleles than non-synonymous mutations. Moreover, we did not include non-synonymous mutations for the phenotypic analysis on the assumption that if they are truly loss of function alleles, sequencing more individuals will eventually yield a nonsense or essential splice allele in those genes.

Phenotyping

Zebrafish were maintained in accordance with UK Home Office regulations, UK Animals (Scientific Procedures) Act 1986, under project licence 80/2192, which was reviewed by The Wellcome Trust Sanger Institute Ethical Review Committee.

Heterozygous F2 fish were randomly incrossed and upon egg collection F2 adults were fin clipped and kept as isolated breeding pairs. For each family we aimed to phenotype 12 pairs, over 3 weeks of breeding. Each clutch of eggs, which was labelled with the breeding pair ID, was sorted into three 10cm petri dishes of ~50 embryos each. Embryos were incubated at 28.5°C. Previous mutagenesis screens were used as a reference for the phenotyping [27,28]. Those phenotypes studied were: day 1 – early patterning defects, early arrest, notochord, eye development, somites, patterning and cell death in the brain; day 2 – cardiac defects, circulation of the blood, pigment (melanocytes), eye and brain development; day 3 – cardiac defects, circulation of the blood, pigment (melanocytes), movement and hatching; day 4 – cardiac defects, movement, pigment (melanocytes) and muscle defects; day 5 – behaviour (hearing, balance, response to touch), swim bladder, pigment (melanocytes, xanthophores and iridophores), distribution of pigment, jaw, skull, axis length, body shape, notochord degeneration, digestive organs (intestinal folds, liver and pancreas), left-right patterning. In the first round of the phenotyping, all phenotypic embryos were discarded. At 5 dpf, >48 phenotypically wild-type embryos were harvested. Embryos were fixed in 100% methanol and stored at −20°C until genotyping was initiated. In the second round, F2s that were heterozygous for a suspected causal mutation were re-crossed. All phenotypes observed in those clutches of embryos were counted, documented and photographed. Phenotypic embryos were fixed in 100% methanol and at 5 dpf 48 phenotypically wild-type embryos were also collected. The first round genotyping results were assessed using a Chi-squared test with a p-value cut off of <0.05. If the number of homozygous embryos was above the cut-off (i.e. in the expected 25% ratio), the allele was deemed to not cause a phenotype within the first 5 dpf. If the number of homozygous embryos was below the cut-off, the allele was carried forward into the second round of phenotyping. In the second round, we aimed to genotype 48 embryos for each phenotype, ideally from multiple clutches. An allele was documented as causing a phenotype if the phenotypic embryos were homozygous for the allele. We allowed up to 10% of embryos for a given phenotype to not be homozygous, to account for errors in egg collection. Such alleles were outcrossed for further genotyping with F4 embryos at a later date. Where possible, alleles were also submitted to complementation tests.

Genotyping

Embryo DNA was prepared by first removing the embryos from 100% methanol and individual embryos were placed into wells of a 96-well plate. The well positions of phenotypic and non-phenotypic embryos were documented. Incubating the plate at 80°C for 15 min evaporates all remaining methanol. DNA was extracted from embryos by incubating in 25 μl of lysis buffer (25 mM NaOH + 0.2 mM EDTA) at 95°C for 30 min. Following this, a volume of 25 μl of neutralisation buffer (40 mM Tris-HCl) was added. Genotyping was performed using the competitive allele-specific PCR (KASP) genotyping system (KBioscience). Fin clip and embryo DNA was diluted to a working concentration of 1.25-12.5 ng/μl. A volume of 4 μl of DNA was pipetted into black 384 well hard shell PCR plates and dried down at room temperature. When the genotyping was performed, the DNA was resuspended by adding a 4 μl PCR mix, according to the manufacturer’s protocol (KBioscience). Genotyping PCR results were analysed using a PHERAstar plus (BMG labtech) and the software KlusterCaller (KBioscience).

Cryopreservation of alleles

Sperm from individual males was collected by abdominal massage into 10 μl glass capillaries. The sperm was expelled into 245 μl 10% N,N-dimethylacetamide (DMA) in BSMIS (75 mM NaCl, 70 mM KCl, 2 mM $CaCl_2$, 1 mM $MgSO_4$, 20 mM Tris pH 8.0) and mixed briefly by pipetting up and down. Eight aliquots of 35 μl each were pipetted directly into 2 ml cryovials and immediately transferred into a pre-chilled 50 ml Falcon tube on a dry ice/ ethanol bath. After 30 min the samples were moved into liquid nitrogen for long-term storage. Storage location, date of collection and sample quality were documented together with genotyping data for each archived male. Representative sperm samples were tested by *in vitro* fertilisation.

For *in vitro* fertilisation sperm samples were thawed by addition of 500 μl 37°C BSMIS to the frozen specimen. The sperm was then activated by the addition of 500 μl 28°C 0.5% fructose in egg water (0.018% (w/v) synthetic sea salt in RO water) and immediately mixed with fresh eggs in a 6cm glass Petri dish. Eggs were obtained by squeezing females and used within a few minutes. Sperm motility and egg quality were monitored and documented.

After about 50 seconds the glass dish was filled with egg water and eggs were transferred into a 10 cm plastic Petri dish. Fertility rates were checked after incubating the eggs for a few hours at 28.5°C.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Pete Ellis, Eleanor Markham Henk van Roekel, Pim Toonen, Jose van de Belt, Josine Mudde and Sara Widaa for technical assistance, and Barbara Novak, Yong Yi, and Emily LeProust from Agilent Technologies. We thank everyone at The Zebrafish International Resource Center, Eugene, Oregon and the European Zebrafish Resource Center, Karlsruhe, Germany for stocking and distributing alleles. We thank members of the Wellcome Trust Sanger Institute RSF and DNA pipelines. We also thank Gareth Powell, John Collins and Florence L. Marlow for critical reading of the manuscript. This work was funded through a core grant to the Sanger Institute by the Wellcome Trust grant number 098051, the US National Institutes of Health (5R01HG004819), the EU Sixth Framework Programme (ZF-MODELS, contract number LSHG-CT-2003-503496), and the EU Seventh Framework Programme (ZF-HEALTH). F.v.E. is supported by the UK Medical Research Council (MRC) and E.C. is supported by the SmartMix program (SSM06010**)** from the Dutch government.

References

1. Gossler A, Joyner AL, Rossant J, Skarnes WC. Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. Science. 1989; 244:463–465. [PubMed: 2497519]

2. Skarnes WC, Auerbach BA, Joyner AL. A gene trap approach in mouse embryonic stem cells: the lacZ reported is activated by splicing, reflects endogenous gene expression, and is mutagenic in mice. Genes Dev. 1992; 6:903–918. [PubMed: 1592261]

3. Ringwald M, et al. The IKMC web portal: a central point of entry to data and resources from the International Knockout Mouse Consortium. Nucleic Acids Res. 2011; 39:D849–855. doi:10.1093/nar/gkq879. [PubMed: 20929875]

4. Church DM, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009; 7:e1000112. doi:10.1371/journal.pbio.1000112. [PubMed: 19468303]

5. Waterston RH, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420:520–562. doi:10.1038/nature01262. [PubMed: 12466850]

6. Zhang J, Talbot WS, Schier AF. Positional cloning identifies zebrafish one-eyed pinhead as a permissive EGF-related ligand required during gastrulation. Cell. 1998; 92:241–251. [PubMed: 9458048]

7. Golling G, et al. Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. Nat Genet. 2002; 31:135–140. doi:10.1038/ng896. [PubMed: 12006978]

8. Amsterdam A, et al. Identification of 315 genes essential for early zebrafish development. Proc Natl Acad Sci U S A. 2004; 101:12792–12797. doi:10.1073/pnas.0403929101. [PubMed: 15256591]

9. Sun Z, et al. A genetic screen in zebrafish identifies cilia genes as a principal cause of cystic kidney. Development. 2004; 131:4085–4093. doi:10.1242/dev.01240. [PubMed: 15269167]

10. Nasevicius A, Ekker SC. Effective targeted gene ‘knockdown’ in zebrafish. Nat Genet. 2000; 26:216–220. doi:10.1038/79951. [PubMed: 11017081]

11. Kettleborough RN, Bruijn E, Eeden F, Cuppen E, Stemple DL. High-throughput target-selected gene inactivation in zebrafish. Methods Cell Biol. 2011; 104:121–127. doi:10.1016/B978-0-12-374814-0.00006-9. [PubMed: 21924159]

12. Sood R, et al. Methods for reverse genetic screening in zebrafish by resequencing and TILLING. Methods. 2006; 39:220–227. doi:10.1016/j.ymeth.2006.04.012. [PubMed: 16828311]

13. Wienholds E, et al. Efficient target-selected mutagenesis in zebrafish. Genome Res. 2003; 13:2700–2707. doi:10.1101/gr.1725103. [PubMed: 14613981]

14. Meng X, Noyes MB, Zhu LJ, Lawson ND, Wolfe SA. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. Nat Biotechnol. 2008; 26:695–701. doi:10.1038/nbt1398. [PubMed: 18500337]

15. Doyon Y, et al. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. Nat Biotechnol. 2008; 26:702–708. doi:10.1038/nbt1409. [PubMed: 18500334]

16. Huang P, et al. Heritable gene targeting in zebrafish using customized TALENs. Nat Biotechnol. 2011; 29:699–700. doi:10.1038/nbt.1939. [PubMed: 21822242]

17. Sander JD, et al. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. Nat Biotechnol. 2011; 29:697–698. doi:10.1038/nbt.1934. [PubMed: 21822241]

18. Howe, K., et al. The Zebrafish Reference Genome Sequence and its Relationship to the Human Genome. 2012. submission

19. Collins JE, White S, Searle SM, Stemple DL. Incorporating RNA-seq data into the Zebrafish Ensembl Gene Build. Genome Res. 2012 doi:10.1101/gr.137901.112.

20. Stemple DL. TILLING--a high-throughput harvest for functional genomics. Nat Rev Genet. 2004; 5:145–150. doi:10.1038/nrg1273. [PubMed: 14726927]

21. Consortium, T. G. P. A map of human genome variation from population-scale sequencing. Nature. 2011; 467:1061–1073.

22. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. doi:10.1038/nature11632. [PubMed: 23128226]

23. Ayadi A, et al. Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. Mamm Genome. 2012; 23:600–610. doi:10.1007/s00335-012-9418-y. [PubMed: 22961258]

24. Seiler C, et al. Myosin VI is required for structural integrity of the apical surface of sensory hair cells in zebrafish. Dev Biol. 2004; 272:328–338. doi:10.1016/j.ydbio.2004.05.004. [PubMed: 15282151]

25. Manfroid I, et al. Zebrafish sox9b is crucial for hepatopancreatic duct development and pancreatic endocrine cell regeneration. Dev Biol. 2012; 366:268–278. doi:10.1016/j.ydbio.2012.04.002. [PubMed: 22537488]

26. Steffen LS, et al. The zebrafish runzel muscular dystrophy is linked to the titin gene. Dev Biol. 2007; 309:180–192. doi:10.1016/j.ydbio.2007.06.015. [PubMed: 17678642]

27. Haffter P, et al. The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. Development. 1996; 123:1–36. [PubMed: 9007226]

28. Driever W, et al. Development. 1996; Vol. 123:37–46. [PubMed: 9007227]

29. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011; 12:443–451. doi:10.1038/nrg2986. [PubMed: 21587300]

30. McLaren W, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010; 26:2069–2070. doi:10.1093/bioinformatics/btq330. [PubMed: 20562413]

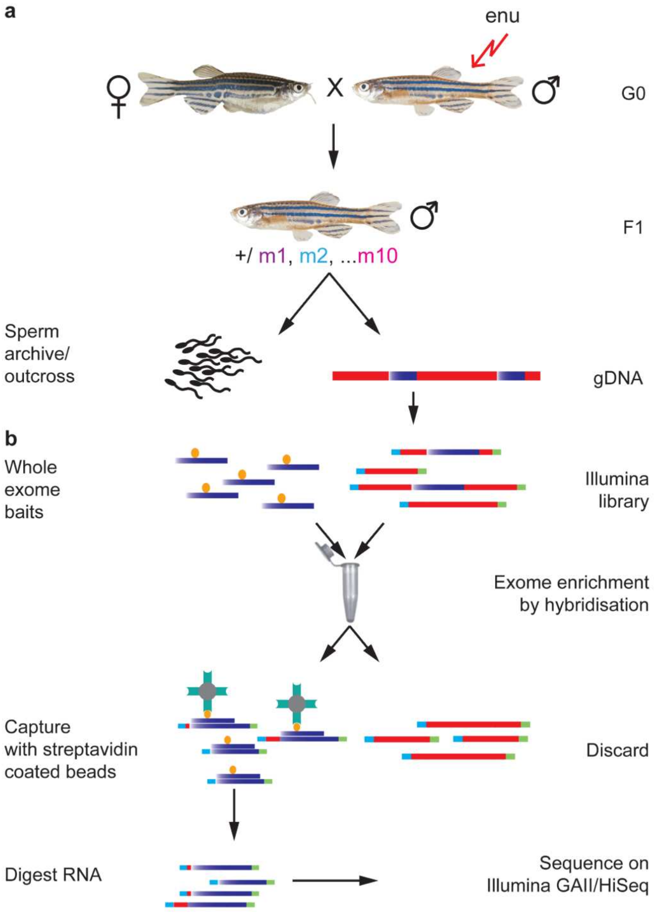

Figure 1. Exome sequencing

a, ENU-mutagenised G0 males are outcrossed to create a population of F1s. Genomic DNA is taken from F1s, which are either outcrossed or cryopreserved as sperm samples. **b,** F1 genomic DNA is then subjected to exome sequencing. Illumina libraries are made and hybridised to the 120mer biotinylated RNA whole exome baits. Streptavidin coated magnetic beads capture genomic DNA hybridised to the RNA baits, and all other DNA is discarded. Exome-enriched DNA fragments are sequenced. Blue represents exonic and red non-coding genomic DNA.

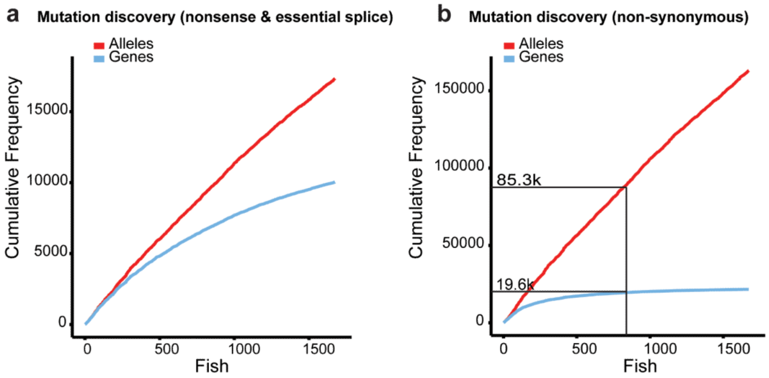

Figure 2. Mutation detection

a, The cumulative detection of nonsense and essential splice alleles. As each mutagenised library displayed different rates of mutagenesis the order that exomes were sequenced was randomised. **b,** The detection of non-synonymous mutations. Sequencing 808 exomes resulted in the identification of 85,338 non-synonymous alleles in 19655 genes corresponding to 75% of all protein-coding genes.

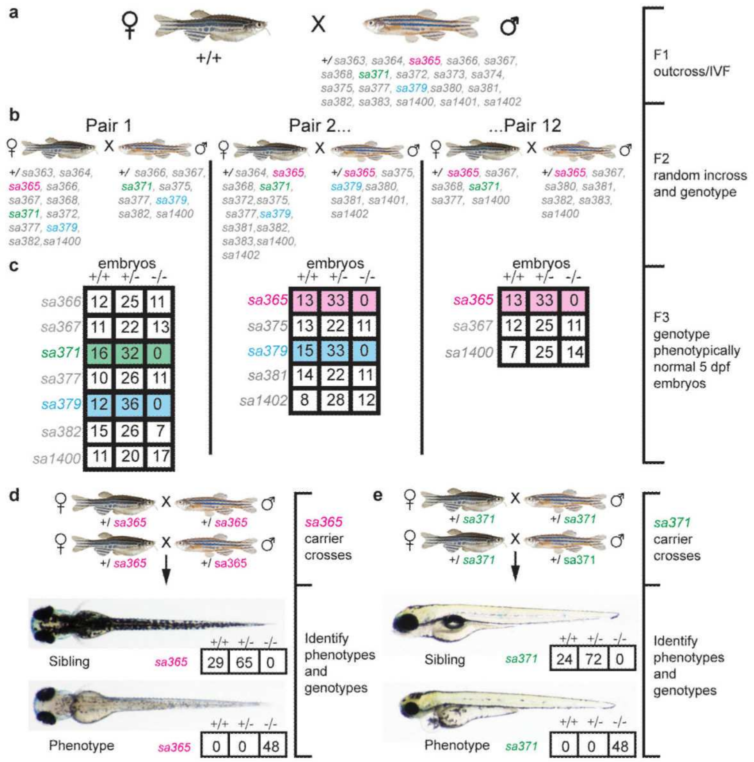
a
♀ +/+
X
♂ +/ sa363, sa364, sa365, sa366, sa367, sa368, sa371, sa372, sa373, sa374, sa375, sa377, sa379,sa380, sa381, sa382, sa383, sa1400, sa1401, sa1402
F1 outcross/IVF
b
Pair 1
♀ +/ sa363, sa364, sa365, sa366, sa367, sa368, sa371, sa372, sa377, sa379, sa382,sa1400
X
♂ +/ sa366, sa367, sa371, sa375, sa377, sa379, sa382, sa1400
Pair 2...
♀ +/ sa364, sa365, sa368, sa371, sa372,sa375, sa377,sa379, sa381,sa382, sa383,sa1400, sa1402
X
♂ +/ sa365, sa375, sa379,sa380, sa381, sa1401, sa1402
...Pair 12
♀ +/ sa365, sa367, sa368, sa371, sa377, sa1400
X
♂ +/ sa365, sa367, sa380, sa381, sa382, sa383, sa1400
F2 random incross and genotype
c
embryos
+/+ +/- -/-
sa366 12 25 11
sa367 11 22 13
sa371 16 32 0
sa377 10 26 11
sa379 12 36 0
sa382 15 26 7
sa1400 11 20 17
embryos
+/+ +/- -/-
sa365 13 33 0
sa375 13 22 11
sa379 15 33 0
sa381 14 22 11
sa1402 8 28 12
embryos
+/+ +/- -/-
sa365 13 33 0
sa367 12 25 11
sa1400 7 25 14
F3 genotype phenotypically normal 5 dpf embryos
d
♀ +/ sa365 X ♂ +/ sa365
♀ +/ sa365 X ♂ +/ sa365
sa365 carrier crosses
Sibling sa365
+/+ +/- -/-
29 65 0
Phenotype sa365
+/+ +/- -/-
0 0 48
Identify phenotypes and genotypes
e
♀ +/ sa371 X ♂ +/ sa371
♀ +/ sa371 X ♂ +/ sa371
sa371 carrier crosses
Sibling sa371
+/+ +/- -/-
24 72 0
Phenotype sa371
+/+ +/- -/-
0 0 48
Identify phenotypes and genotypes

Figure 3. Phenotypic analysis of alleles

a, F1 individuals were outcrossed to produce an F2 family. The induced disruptive alleles for one family are shown. **b,** F2s were incrossed and genotyped. **c,** First round, phenotypically wild-type embryos were collected from each clutch at 5 dpf and genotyped for the mutations heterozygous in both parents. The number of homozygous mutant F3 embryos was assessed using a Chi-squared test (p-value cut-off <0.05). Mutations homozygous in less than 25% of embryos were suspected to cause a phenotype. Here, there were no homozygous embryos in the phenotypically wild-type set for the alleles *sa365*, *sa371* and *sa379*. **d-e,** Second round, phenotypes present within each incross were genotyped for putative phenotypic mutations. **d,** *slc22a7b*^{sa365} shows pigment phenotype. **e,** *mphosph10*^{sa371} shows small head and pericardiac oedema phenotype. *lamc1*^{sa379} mutants are shown in Fig. 4 l, n.

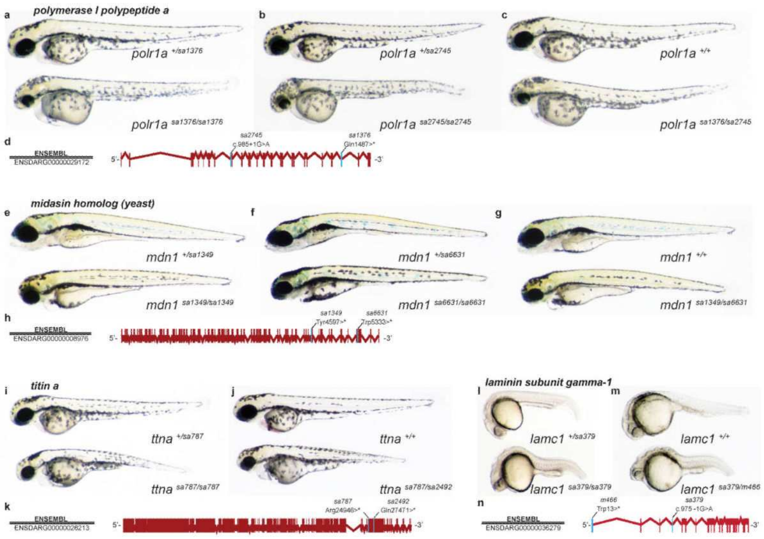

Figure 4. Confirmation of causality through complementation crosses

Depicted are four examples, *polymerase I polypeptide a (polr1a)* (**a-d**), *midasin homologue (yeast) (mdn1)* (**e-h**), *titin a (ttna)* (**i-k**) and *laminin subunit gamma-1 (lamc1)* (**l-n**), where heterozygous carriers of two independent alleles in the same gene were used to generate compound heterozygote offspring. Where possible incrosses of individual alleles are shown as well. In all images non-phenotypic siblings are above and phenotypic homozygous mutant or compound heterozygous embryos below. **a-c,** At 48 hpf embryos homozygous for either *sa1376*, *sa2745* or compound heterozygous for *sa1376* and *sa2745* have small eyes, a hydrocephalic hindbrain and pericardiac oedema. **d,** *sa2745* disrupts a splice donor site through a G>A transition at the first intronic nucleotide 3′ of coding nucleotide 985. Allele *sa1376* is a C>T transition producing a premature STOP codon at amino acid (aa) 1487. **e-g,** At 96 hpf embryos homozygous for either *sa1349*, *sa6631* or compound heterozygous for *sa1349* and *sa6631* have smaller heads with malformed jaws and mild pericardiac oedema. **h,** *sa1349* and *sa6631* produce premature STOP codons at aa 4597 (T>A transversion) and aa 5333 (G>A transition), respectively. **i, j,** At 48 hpf embryos homozygous for *sa787* or compound heterozygous for *sa787* and *sa2492* are growth retarded, paralysed and have pericardiac oedema. **k,** Alleles *sa787* and *sa2492* produce premature STOP codons at aa 24946 (C>T transition) and aa 27471 (C>T transition), respectively. **l, m,** At 24 hpf embryos homozygous for *sa379* or compound heterozygous for *sa379* and *m466* are shorter with an undifferentiated notochord, and brain and eye malformations. **n,** Allele *m466* is a G>A transition producing a premature STOP codon at aa 13. Allele *sa379* disrupts a splice acceptor site through a G>A transition one nucleotide 5′ of coding nucleotide 975.