Title

The Vertebrate Codex Gene Breaking Protein Trap Library For Genomic Discovery and Disease Modeling Applications

Authors

Noriko Ichino[1], MaKayla Serres[1], Rhianna Urban[1], Mark Urban[1], Kyle Schaefbauer[1], Lauren Greif[1], Gaurav K. Varshney[3], Kimberly J. Skuster[1], Melissa McNulty[1], Camden Daby[1], Ying Wang[4], Hsin-kai Liao[4], Suzan El-Rass[5], Yonghe Ding[1,2], Weibin Liu[1,2], Lisa A. Schimmenti[1], Sridhar Sivasubbu[6], Darius Balciunas[7], Matthias Hammerschmidt[8], Steven A. Farber[9], Xiao-Yan Wen[5], Xiaolei Xu[1,2], Maura McGrail[4], Jeffrey J. Essner[4], Shawn Burgess[10], Karl J. Clark[1*], Stephen C. Ekker[1*]

*Corresponding authors

Author contributions

SCE and KJC conceived research; NI, SCE, KJC for basic experimental design including detailed analyses of the collection; MS, CD, MU, SER, YD, RU, MM, JJE, XX, DB, SCE and KJC generated GBT collection, NI, KS, MS, LG, CD, MU, RU, YD, SAF, WL and KJC conducted phenotype screening of GBT mutant lines; NI, KS, MS, LG, CD, MU, RU, YD, WL and KJC conducted molecular biology analyses; GV and SB conducted next generation sequencing; NI, KS, SCE and KJC conducted bioinformatics-based analyses; LAS, NI, KJC and SCE conducted comparative genomics analyses, NI, KS, MS, LG, KJS and SCE wrote the manuscript; SCE, KJC, JJE, XX, MH, SAF, XYW, SB, XX and SL consulted to this research.

Funding: Supported by NIH grants (GM63904, DA14546, DK093399 and HG 006431), Natural Sciences and Engineering Research Council of Canada, grant RGPIN 05389-14 and the Mayo Foundation.

Affiliations

1. Department of Biochemistry and Molecular Biology, Mayo Clinic, Minnesota, USA

2. Department of Cardiovascular Medicine, Mayo Clinic, Minnesota, USA

3. Functional & Chemical Genomics Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, 73112, USA

4. Department of Genetics, Development and Cell Biology, Iowa State University, Iowa, USA

5. Zebrafish Centre for Advanced Drug Discovery & Keenan Research Centre for Biomedical Science, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, Ontario, Canada.

6. CSIR–Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India.

7. Department of Biology Temple University, 435 Biology- Life Sciences Building, 1900 North 12th Street, Philadelphia, PA 19122, USA

8. Institute for Developmental Biology, Cologne University, Köln, Germany

9. Carnegie Institution, Baltimore, Maryland, USA

10. Translational and Functional Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892-8004, USA.

Keywords

1 Gene-break transposon, zebrafish, disease model, human genetic disorders, gene ontology, Lightsheet

2 microscopy

Abstract

The zebrafish is a powerful model to explore the molecular genetics and expression of the vertebrate genome. The gene break transposon (GBT) is a unique insertional mutagen that reports the expression of the tagged member of the proteome while generating Cre-revertible genetic alleles. This 1000+ locus collection represents novel codex expression data from the illuminated mRFP protein trap, with 36% and 87% of the cloned lines showcasing to our knowledge the first described expression of these genes at day 2 and day 4 of development, respectively. Analyses of 183 molecularly characterized loci indicate a rich mix of genes involved in diverse cellular processes from cell signaling to DNA repair. The mutagenicity of the GBT cassette is very high as assessed using both forward and reverse genetic approaches. Sampling over 150 lines for visible phenotypes after 5dpf shows a similar rate of discovery of embryonic phenotypes as ENU and retroviral mutagenesis. Furthermore, five cloned insertions were in loci with previously described phenotypes; embryos homozygous for each of the corresponding GBT alleles displayed strong loss of function phenotypes comparable to published mutants using other mutagenesis strategies (*ryr1b*, *fras1*, *tnnt2a*, *edar* and *hmcn1*). Using molecular assessment after positional cloning, to date nearly all alleles cause at least a 99+% knockdown of the tagged gene. Interestingly, over 35% of the cloned loci represent 68 mutants in zebrafish orthologs of human disease loci, including nervous, cardiovascular, endocrine, digestive, musculoskeletal, immune and integument systems. The GBT protein trapping system enabled the construction of a comprehensive protein codex including novel expression annotation, identifying new functional roles of the vertebrate genome and generating a diverse collection of potential models of human disease.

Introduction

With the generation of more than 100 sequenced vertebrate genomes (Meadows & Lindblad-Toh, 2017), the current key question is how to determine the role(s) of uncharacterized gene products in specific biological and pathological processes. For example, genes associated with human disease are being discovered at a rapid rate. However, the biological functions underlying this linkage is often unclear (Kettleborough et al., 2013). Model system science using loss of function approaches has been essential to the annotation of the genome to date including the discovery of novel processes and the biological mechanisms underlying disease (Stoeger, Gerlach, Morimoto, & Nunes Amaral, 2018).

Among vertebrates, *Danio rerio* (zebrafish) has emerged as an outstanding model organism amenable to both forward and reverse genetic approaches. In addition, the natural transparency of the zebrafish embryo and larvae enables the unprecedented ability to non-invasively collect a rich set of expression data for the proteome and in the context of an entire living vertebrate. We describe here a 1000+ collection of zebrafish lines made using the Protein Trap Gene- Breaking Transposon (GBT;(Clark, Balciunas, et al., 2011)to develop such a codex for the comparative vertebrate genomics field (Meadows & Lindblad-Toh, 2017), (Clark, Balciunas, et al., 2011).

The initial pGBT-RP 2.1 (RP2.1) vector has several features that efficiency cooperate to report gene sequence, expression and function (Clark, Balciunas, et al., 2011). Two main reporter components include a 5’ protein trap and a 3’ exon trap, with the entire cassette flanked by inverted terminal repeats (ITR) of the mini*Tol2* transposon to effectively deliver the transgene as single copy integrations into the zebrafish genome. In cases where RP2 integrates in the sense orientation of a transcription unit, the protein trap’s splice acceptor overrides normal splicing of the transcription unit, creating a fusion between endogenous upstream exons and the monomeric RFP (mRFP) reporter sequences. The protein-trap domain in RP2.1 generates the expression profile, including subsequent protein localization and

accumulation when a functional in-frame fusion between the start codon-deficient mRFP reporter and the tagged protein. Mutagenesis is accomplished by the strong internal polyadenylation and putative border element, effectively truncating the endogenously tagged protein. The GBT mutagenesis system represented the first step toward a 'codex' of protein expression and functional annotation of the vertebrate genome (Clark, Balciunas, et al., 2011).

We report here the development of a series of GBT protein trap vectors including versions to trap expression in each of the three potential reading frames. In addition, we modified the 3' exon trap to use a localized BFP rather than the more commonly used GFP to more effectively use these lines in conjunction with other transgenic fish. We deployed these vectors at scale, generating over 1000 protein trap lines with visible mRFP expression at either 2dpf (end of embryogenesis) or 4dpf (larval stage), with 36% and 87% of the cloned lines showcasing to our knowledge the first described expression of these genes at these stages, respectively. We used forward and reverse genetic tests to assess the mutagenicity of these vectors, noting similar rates of visible phenotypes at 5dpf as ENU and retroviral screening tools. We re-isolated five previously described loci, and embryos homozygous for each of the corresponding GBT alleles displayed strong loss of function phenotypes comparable to these previously published mutants generated using other mutagenesis strategies (*ryr1b*, *fras1*, *tnnt2a*, *edar* and *hmcn1*). Molecular assessment after positional cloning shows that nearly all alleles cause at least a 99+% knockdown of the tagged gene. Interestingly, over 35% of the cloned loci represent 68 mutants in zebrafish orthologs of human disease loci, including nervous, cardiovascular, endocrine, digestive, musculoskeletal, immune and integument systems. The GBT protein trapping system enabled the construction of a comprehensive protein codex including novel expression annotation, identifying new functional roles of the vertebrate genome and generating a diverse collection of potential models of human disease.

Materials and Methods

Zebrafish husbandry

All zebrafish (*Danio rerio*) was maintained according to the guidelines and the standard procedures established by the Mayo Clinic Institutional Animal Care and Use Committee (Mayo IACUC). The Mayo IACUC approved all protocols involving live vertebrate animals (A23107, A21710 and A34513).

Generating GBT constructs, RP2 and RP8 series

pGBT-RP8.2 and -RP8.3 were made by combining three restriction endonuclease fragments of pGBT-RP8.1, a 2.2 kb AflII to AgeI, a 0.7 kb EcoRI to SpeI, and a 3.0kb SpeI to AflII, with a short adapter to close the space between AgeI and EcoRI that effectively removed one or two thymine nucleotides just following the splice acceptor prior to the AUG-less mRFP cassette. For pGBT-RP8.2, Adapter-GBT(+2) was made by annealing oligos adapter-GBT(+2)-a [CCGGTTTTCTCATTCATTTACAGTCAGCCGG] and adapter-GBT (+2)-b [AATTCCGGCTGACTGTAAATGAATGAGAAAA]. For pGBT-RP8.3, Adapter-GBT(+3) was made by annealing oligos adapter-GBT (+3)-a [CCGGTTTTCTCATTCATTTACAGCAGCCGG] and adapter-GBT(+3)-b [AATTCCGGCTGCTGTAAATGAATGAGAAAA].

pGBT-RP2.2 and -RP2.3 were made by combining three restriction endonuclease fragments of pGBT-RP2.1 (Clark, Balciunas, et al., 2011), a 3.6kb BlpI to AgeI, a 1.9kb EcoRI to AvrII, and a 3.55kb AvrII to BlpI, with a short adapter to close the space between AgeI and EcoRI that effectively removed one or two thymine nucleotides just following the splice acceptor prior to the AUG-less mRFP cassette. For pGBT-RP2.2, Adapter-GBT(+2) was made by annealing oligos adapter-GBT(+2)-a and adapter-GBT (+2)-b . For pGBT-RP2.3, Adapter-GBT(+3) was made by annealing oligos adapter-GBT (+3)-a and adapter-GBT(+3)-b.

pGBT-RP8.1 was made by cloning a mini-intron derived from carp beta actin intron 1 into pGBT-RP7.1. The 234bp SalI to XhoI mini-intron fragment was isolated from pCR4-bactmIntron following digestion. The pGBT-RP7.1 plasmid was digested with XhoI so that the SalI to XhoI fragment was cloned between the gamma-crystallin promoter and nls tagBFP.

pCR4-bactmIntron was made by removing a 1.1kb internal portion of the carp beta actin intron 1 by digestion of pCR4-bact_I1 with BstBI and BssHII, followed by filling in 5' overhangs and ligating remaining vector fragment.

pCR4-bact_I1 was cloning a PCR product containing the carp beta-actin intron into pCR4-TOPO (Invitrogen). The intron was amplified from pGBT-RP2.1 (Clark, Balciunas, et al., 2011) using MISC-bact_exon-F1 [CAGCTAGTGCGGAATATCATCTGCC] and MISC-bact_intron-R1 [CTTCTCGAGGTGAATTCCGGCTGAACTGTA] primers.

pGBT-RP7.1 was made by replacing a 501bp PstI to PstI fragment of pGBT-RP6.1 with a 480bp PstI to PstI fragment of pRP2.1. This changed the nucleotide sequence between the carp beta-actin splice acceptor to replicate the sequences in pGBT-RP2.1. pGBT-RP7.1 was never directly tested in zebrafish.

pGBT-RP6.1 was made by flipping the internal trap cassette relative the Tol2 inverted terminal repeats in pGBT-RP5.1. To do this, pGBT-RP5.1 was cut with EcoRV and SmaI. The 2.27kb EcoRV to SmaI vector backbone fragment, which included the ITRs, was ligated to the 3.51kb EcoRV to SmaI trap

fragment. pGBT-RP6.1 was then selected based on the right ITR of Tol2 being in front of the RFP trap, which is the same orientation of pGBT-RP2.1.

pGBT-RP5.1 was made by cloning a PCR product with the AUG-less mRFP into pre(-1)GBT-RP5.1. The 698bp mRFP* PCR product was obtained by amplification of pGBT-R15 (Clark, Balciunas, et al., 2011) with CDS-mRFP*-F1 [AAGAATTCGAAGGTGCCTCCTCCGAGGATGTCATCAAGG] and CDS-mRFP-R1 [AAACTAGTCTTAGGCTCCGGTGGAGTGGCGG]. Prior to cloning the PCR mRFP* product was digested with EcoRI and SpeI to prepare the ends for subcloning into pre(-1)GBT-RP5.1 that was opened between the carp beta actin splice acceptor and the ocean pout terminator.

pre(-1)GBT-RP5.1 was made by cloning 1.2kb SpeI to AvrII fragment from pGBT-PX (Sivasubbu et al., 2006) that contained the ocean pout terminator into the SpeI site of pre(-2)GBT-RP5.1. The resulting products were screened for the proper orientation of the ocean pout terminator relative to the carp beta actin splice acceptor.

pre(-2)GBT-RP5.1 was made by inserting an expression cassette to make a 3' poly(A) trap that makes blue lenses. A 1.15kb SpeI to BglII fragment from pKTol2gC-nlsTagBFP was cloned into pre(-3)GBT-RP5.1 that had been cut with AvrII and BglII. This moved the *Xenopus* gamma crystallin promoter driving a nuclear-localized TagBFP in front of the carp beta actin splice donor within pre(-3)GBT-RP5.1 to create a localized BFP poly(A) trap signal replacing the ubiquitous GFP signal that was in pGBT-RP2.1.

pre(-3)GBT-RP5.1 was made by cloning a 492bp XmaI to NheI scaffold fragment from pUC57-I-SceI_loxP_splice into pKTol2-SE (Clark, Balciunas, et al., 2011) opened with XmaI and NheI. pUC57-I-SceI_LoxP_Splice contains a synthetic sequence (see below) cloned into pUC57 (Genscript). The scaffold contains an I-SceI site; loxP site; carp beta actin splice acceptor; cloning sites for mRFP, ocean pout terminator, and BFP lens cassettes; carp beta actin splice donor; loxP site; and an I-SceI site.

[cccgggatagggataacagggtaatataacttcgtatagcatacattatacgaagttatcgttaccacccactagcggtcagactgcagattgcagcac
gaaacaggaagctgactccacatggtcacatgctcactgaagtgttgacttccctgacagctgtgcactttctaaaccggttttctcattcatttacagttca
gcctgttacctgcactcaccgacaagctgttaccctggaattcgtttaaacactagtcaccggcgttcctaggttataagatctacctaaggtgagttgatct
ttaagctttttacattttcagctcgcatatatcaattcgaacgtttaattagaatgtttaaataaagctagattaaatgattaggctcagttaccggtctttttttct
catttacactgagctcaagacgtctgataacttcgtatagcatacattatacgaagttattaccctgttatccctatggctagc]

Generating GBT collection

Generation of the GBT collection was based on the prior described protocols (Clark, Balciunas, et al., 2011; Clark, Urban, Skuster, & Ekker, 2011; J. Ni et al., 2016).

Fluorescent microscopy of mRFP reporter protein expression

Larvae were treated with 0.2 mM phenylthiocarbamide at 1 dpf to inhibit pigment formation. The anesthetized fish were mounted in 1.5% agarose (Fisher Scientific BP1360) prepared with 0.017mg/ml tricaine solution in an agarose column in the imaging chamber. The protocol of ApoTome microscopy was described in previous publication. (Clark, Balciunas, et al., 2011) For Lightsheet microscopy, larval zebrafish were anesthetized with 0.017g/ml tricaine (Ethyl 3-aminobenzoate methanesulfonate salt) in embryo water during imaging procedure. To capture RFP expression patterns of 2 dpf and 4 dpf larval zebrafish, LP 560 nm filter as excitation and LP 585nm as emission was used for Lightsheet microscopy.

The sagittal-, dorsal-, and ventral- oriented z-stacks of the mRFP expression were captured at either 50x magnification using an ApoTome microscope (Zeiss) with a 5x/0.25 NA dry objective (Zeiss) or 50x magnification using a Lightsheet Z.1 microscope (Zeiss) 5x/0.16 NA dry objective. Each set of images were obtained from the same larva and the images shown are composites of the maximum image projections of the z-stacks obtained from each direction.

Sperm Cryopreservation

Sperm collection and cryopreservation was initially based on the protocol described in (Draper & Moens, 2009) and moved to the ZIRC protocol described in (Matthews et al., 2018).

Genomic DNA isolation

Genomic DNA was isolated from F1 fish tail biopsies to conduct next generation sequencing and from both WT and heterozygous larva to manually perform the PCR-based mRFP linkage analysis. Zebrafish larva were individually placed to 0.2 ml PCR tubes and sacrificed to extract genomic DNA in 50 mM NaOH for 20 min at 95 C^{o}.

PCR-based linkage analysis of GBT insertions loci

TALE-PCR: The protocol used was designed to amplify and clone junction fragments from Tol2-based gene-break transposons (GBT) in the zebrafish genome. Although modified, it is based on a protocols received from Alexi Parnov, Vladimir Korsch, and Karuna Sampath. The following primer mixtures (containing 0.4 µM GBT specific primer and 2 µM DP primer) were prepared: for primary PCR: 5R-mRFP-P1/DP1, 5R-mRFP-P1/DP2, 5R-mRFP-P1/DP3, 5R-mRFP-P1/DP4, 3R-GM2-P1/DP1, 3R-GM2-P1/DP2, 3R-GM2-P1/DP3, 3R-GM2-P1/DP4, 3R-tagBFP-P1/DP1, 3R-tagBFP-P1/DP2, 3R-

tagBFP-P1/DP3, 3R-tagBFP-P1/DP4 ; for secondary PCR: 5R-mRFP-P2/DP1, 5R-mRFP-P2/DP2, 5R-mRFP-P2/DP3, 5R-mRFP-P2/DP4, 3R-GM2-P2/DP1, 3R-GM2-P2/DP2, 3R-GM2-P2/DP3, 3R-GM2-P2/DP4, 3R-tagBFP-P2/DP1, 3R-tagBFP-P2/DP2, 3R-tagBFP-P2/DP3, 3R-tagBFP-P2/DP4; for tertiary PCR: TAIL-bA-SA/DP1, TAIL-bA-SA/DP2, TAIL-bA-SA/DP3, TAIL-bA-SA/DP4, Tol2-ITR(L)-O1/DP1, Tol2-ITR(L)-O1/DP2, Tol2-ITR(L)-O1/DP3, Tol2-ITR(L)-O1/DP4, Tol2-ITR(L)-O3/DP1, Tol2-ITR(L)-O3/DP2, Tol2-ITR(L)-O3/DP3, Tol2-ITR(L)-O3/DP4. A total of 1 µl of primer mixtures were added to PCR reaction (total volume 25 µl). Cycle settings were as follows. Primary: (1) 95°C, 3 min; (2) 95°C, 20 sec; (3) 61°C, 30 sec; (4) 70°C, 3 min; (5) go to "cycle 2" 5 times; (6) 95°C, 20 sec; (7) 25°C, 3 min; (8) ramping 0.3°/sec to 70°C; (9) 70°C, 3 min; (10) 95°C, 20 sec; (11) 61°C, 30 sec; (12) 70°C, 3 min; (13) 95°C, 20 sec; (14) 61°C, 30 sec; (15) 70°C, 3 min; (16) 95°C, 20 sec; (17) 44°C, 1 min; (18) 70°C, 3 min; (19) go to "cycle 10" 15 times; (20) 70°C, 5 min; Soak at 12 °C . A total of 5 µl of the primary reaction was diluted with 95 µl of 10mM Tris-Cl or TE buffers and 1µl of the mixture was added to the secondary reaction. Secondary: (1) 95°C, 2 min (2) 95°C, 20 sec; (3)61°C, 30 sec; (4) 70°C, 3 min; (5) 95°C, 20 sec; (6) 61°C, 30 sec ; (7) 70°C, 3 min; (8) 95°C, 20 sec; (9) 44°C, 1 min; (10) ramping 1.5°/sec to 70°C; (11) 70°C, 3 min; (12) go to "cycle 2" 15times; (13) 70°C, 5 min; Soak at 12°C. A total of 5 µl of the primary reaction was diluted with 95 µl of 10mM Tris-Cl or TE buffers and 1µl of the mixture was added to the tertiary reaction. Tertiary: (1) 95°C, 2 min; (2) 95°C, 20 sec; (3) 44°C, 1 min; (3) ramping 1.5°/sec to 70°C; (4) 70°C, 3 min; (5) go to "cycle 2" 32 times; (6) 70°C, 5 min; Soak at 12°C. Products of the secondary and tertiary reactions were separated by using 1-1.5% agarose gel. The individual bands from the "band shift" pairs were cut from the gel and purified by using QIAquick Gel Extraction Kit (QIAGEN, Germany), and sequenced by using ABI Cycle Sequencing chemistry (PE Applied Biosystems, CA) and an ABI Prism 310 Genetic Analyzer with Data Collection Software (PE Applied Biosystems, Foster City, CA) supplied by the producer.

5’ and 3’ RACE PCRs: The protocol used was designed to use cDNA to amplify and clone junction fragments of Tol2-based gene-break transposons (GBT) in the zebrafish genome. 5' RACE allows PCR amplification of unknown sequence at the 5' end of a cDNA as long as there is enough known sequence within the cDNA to design two antisense primers. Although modified, it is based on the protocol in described in (Clark, Balciunas, et al., 2011). The following primer mixtures were prepared: for primary PCR: 0.20 µM GBT specific primer (5R-mRFP-P1), and a mix of universal 5’ RACE primers 2.5 µM 5R-UP-S and 0.5 µM 5R-UP-L . Secondary reaction: 25 µM GSP (5R-mRFP-P2), and 25 µM universal primer 5R-N1. The reaction mix used was as follows: Primary: (25 µl reaction) Template (RR-cDNA) 2 µl, Bioline 5X myTaq buffer 5 µl, myTaq 0.25 µl, GSP 0.5 µl, Universal 5’ RACE primer mix (URS mix) 2.0 µl, Water 15.25 µl. Secondary: (50 µl reaction) Template (2:100 dilution 1° PCR Reaction), Bioline 5X myTaq buffer 10 µl, myTaq 0.3 µl, GSP-P2 0.9 µl, URS-P2 0.9 µl, Water 35.9 µl. Cycle Settings are as follows. Primary: (1) 95°, 3'; (2) 95°, 30"; (3) 65°, 30" -0.5°/cycle; (4) 70°, 2'; Go To (2) x 15 cycles; (5) 95°, 30"; (6) 57°, 30"; (7) 70°, 2'; Go To (5) x 20 cycles, (8) 70°, 10'; (9) Soak at 12°. Dilute 2µL of the primary PCR reaction with 198µL of 10mM Tris-Cl; 1mM EDTA pH8.0. Secondary: (1) 95° for 3' (2) 95°, 30"; (3) 63°, 30" -0.5°/cycle; (4) 70°, 2'; Go To (2) x 10 cycles; (5) 95°, 30"; (6) 58°, 30"; (7) 70°, 2'; Go To (5) x 25 cycles; (8) 70°, 10'; Soak at 12°. After the PCR reactions finish 20µL of each sample were run on a 1.2% agarose gel. You may also run 10µL of undiluted primary PCR reactions on the same gel; however, most of the time the bands of interest do not all appear until after the nested PCR reactions have been run. The individual bands from the gel were excised and purified by using QIAquick Gel Extraction Kit (QIAGEN, Germany), and sequenced by using ABI Cycle Sequencing chemistry (PE Applied Biosystems, CA) and an ABI Prism 310 Genetic Analyzer with Data Collection Software (PE Applied Biosystems, Foster City, CA) supplied by the producer.

Forward genetic screening with next-generation sequencing

Isolated genomic DNA (300-500ng) was digested with MseI, and BfaI in parallel for 3h at 37°C and heat inactivated for 10 min at 80°C. The digested samples from each enzyme were pooled with prealiquoted barcoded linker in individual wells. The T4 DNA ligase (New England Biolabs, Inc.) was added, and the reaction mix was incubated for 2 h at 16°C. The linker-mediated PCR was performed in two steps. In the first step, PCR was done with one primer specific to the 3'- ITR (5'-GACTTGTGGTCTCGCTGTTCCTTGG-3') and the other primer specific to linker sequences (5'-GTAATACGACTCACTATAGGGC- 3') using the following conditions: 2 min at 95°C, 25 cycles of 15 sec at 95°C, 30 seconds at 55°C and 30 seconds at 72°C. The PCR products were diluted to 1:50 in dH2O, and a second round of PCR was performed using ITR (5'-TCACTTGAGTAAAATTTTTGAGTACTTTTTACACCTC-3') and linker specific (5' -GCGTGGTCGACTGCGCAT-3') nested primers to increase sensitivity and avoid non- specific amplification using the following conditions: 2 min at 95°C, 20 cycles of 15 sec at 95°C, 30 seconds at 58°C and 30 seconds at 72°C. The nested PCR products from each 96-well plate are pooled and processed for Illumina library preparation as per manufacturer's instructions.

Protein classification of the cloned zebrafish genes

The 183 cloned zebrafish genes are classified by using PANTHER (Mi, Muruganujan, & Thomas, 2013). PANTHER provided protein classes of the molecule coded by the cloned zebrafish genes.

Annotating human orthologues of GBT-tagged genes and disease-causing genes

The human orthologues of 192 cloned zebrafish genes were mainly collected by using a data mining tool, ZebrafishMine (Van Slyke et al., 2018) supported by the ZFIN database. In some cases, the candidates

of human orthologues unlisted in ZFIN database were manually searched by using both Ensembl (https://useast.ensembl.org/index.html) and InParanoid8 (http://inparanoid.sbc.su.se/cgi-bin/index.cgi) databases. In parallel, the candidates were manually identified by the result of BLASTP assembled with human proteins and by the result of an online synteny analysis tool, SynFind (https://genomevolution.org/CoGe/SynFind.pl). If the candidate multiply hit in those manual assessments, it was annotated as a human orthologue. The human phenotype data caused by mutations of 68 human orthologues were collected by using another data mining tool, BioMart (http://useast.ensembl.org/biomart/martview/cfe15ead83199a0b7c7997f5a4ce9e6b) supported by Ensembl database.

Finding Disease Models in Vertebrates

Mouse models were found by using both descriptions of animal models in Online Mendelian Inheritance in Man (OMIM; https://www.omim.org/) and in Mouse Genome Informatics (MGI; http://www.informatics.jax.org/humanDisease.shtml). MGI provided the details of mouse models of human disease, such as the number of models have been established. Zebrafish model were also found by using both OMIM and the Zebrafish Information Network (ZFIN; https://zfin.org/). ZFIN provided all data of fish strains listed in this database.

Gene expression profiling of the cloned zebrafish genes

The cloned genes with unpublished expression data were isolated by using “Gene Expression” tool of ZebrafishMine (Van Slyke et al., 2018). In parallel, some published expression data were also manually searched from ZFIN database or in some references. To isolate the gene with the expression localized in the tissues or the organs shows abnormalities in the causing diseases of the human orthologues, the

mRFP reporter expression patterns of the cloned genes 2 and 4 dpf are manually analyzed using zfishbook database (Clark, Argue, Petzold, & Ekker, 2012).

Results

The features of GBT constructs RP2 and RP8 – capturing all three proteomic reading frames

In our previous study, we reported the intronic-based gene-breaking transposons (GBTs) as effective and revertible loss-of function tools for zebrafish (Clark, Balciunas, et al., 2011). The main features of the RP2.1 vector system are as follows (Fig1A): 1) Genetically engineered cargo is flanked by miniTol2 sequences necessary and sufficient for Tol2 transposase-mediated transposition, an efficient transgenesis vector in zebrafish (Kawakami et al., 2004); (Balciunas et al., 2006); (Urasaki, Morvan, & Kawakami, 2006); 2) protein trap that enables *in vivo* expression selection of the vertebrate proteome. The addition of the AUG-free mRFP reporter has yielded an effective protein trap for both organ-specific and subcellular localization of the tagged locus (Petzold et al., 2009); (Clark, Balciunas, et al., 2011); (Liao et al., 2012);(Xu et al., 2012);(Ding et al., 2013); (Westcot et al., 2015); 3) Mutagenic transcriptional terminator. The 5' cassette is a combination of a strong splice-acceptor (SA), poly adenylation signal (pA), and putative border element (red octagon) in conjunction with a start codon (AUG)-free monomeric red fluorescent protein (mRFP) reporter. These elements have been shown to be effective as a transcriptional stop in zebrafish by hijacking endogenous splicing (Sivasubbu et al., 2006). These elements are very effective at inducing a quantitative knockdown in all 26 lines assessed to date using qRT-PCR; 97% or higher knockdown in all lines (Clark, Balciunas, et al., 2011); (Ding et al., 2013); (Ding et al., 2016), this manuscript). The GBT mutagenesis system is thus an effective first step to creating a gene codex simultaneously combining expression and loss of function genetics.

However, some limitations were noted with the initial RP2.1 vector – notably the effective trapping of transcripts without detectable expression of the mRFP reporter. Molecular cloning of these GFP+/RFP- lines demonstrated the fidelity of expression requiring the capture of an appropriate reading frame. RP2.1 was designed to use one main reading frame, and some lines with expression were noted to

include the use of a secondary splice acceptor (data not shown). To maximize genome coverage of this insertional mutagen, we created all three reading frames of the RP2 and RP8 vector series (Fig. 1). 4) 3' exon trap. These vectors also encodes a 3' exon trap with preferential expression following intragenic insertions in zebrafish (Sivasubbu et al., 2006); (Petzold et al., 2009); (Clark, Balciunas, et al., 2011). The function of this cassette complements the obligate, in-frame protein trapping effect and is used for both quality control during mutagenesis and for genotyping of more weakly expressing protein trap alleles (Fig.1A). In the RP2 vector series, the nearly ubiquitous b-actin promoter drives expression of GFP. Expression of integrated GFP becomes detectable between early developmental stages such as seven- to eight-somite-stage and provides bright expression with a good signal-to-noise ratio at 25 hpf (Davidson AE et al., 2003). However, the ubiquitous GFP expression from the 3' exon trap cassettes could interfere with another fluorescent marker system based on GFP labeling in further studies. The RP8 vector series includes all reading frames for the AUG-free mRFP reporter and a new 3' exon trap cassette with expression of tagBFP driven by the lens-specific gamma-crystalline promoter (Fig. 1B). Using the tissue-specific reporter system with BFP is helpful to easily detect F1 founder with weak mRFP expression and to avoid interference with GFP-based multi-labeling purposes when crossed with other GFP-labeled transgenic fish lines. 5) Revertible mutagenic cassette. The flanking loxP sites enable reversion of the tagged locus by Cre – mediated recombination. This facilitates both somatic (Clark, Balciunas, et al., 2011); (Ding et al., 2013) and germline approaches (Petzold et al., 2009).

We generated more than eleven hundred independent lines by using all six constructs of the GBT system (Supplemental Table 1). We conducted an initial screening expression of the mRFP fusion protein and showed that RP2 and RP8 vector series with all reading frames of mRFP reporter protein readily detects the distribution of the fusion proteins expressed from their own promoter in zebrafish (Supplemental Figure 1).

Annotation of protein localization and trafficking of the GBT strain collection

The ability to non-invasively obtain temporal and spatial expression pattern information is a key feature of these protein trap strains. Pilot data from our first RFP lines rapidly demonstrated this step was going to be a major bottleneck for our pipeline if we used standard documentation methods. Consequently, we established a capillary-based confinement and imaging protocol (SCORE imaging; (Petzold et al., 2010) for quickly capturing and holding living zebrafish for rapid, high quality fluorescent expression in precise longitudinal imaging angles. The animals are easily inserted into a capillary of the correct diameter for the particular fish stage, then placed on the microscope and covered with a matched solution to remove the optical distortion from the capillary housing. The capillary is simply rotated for precise 0 degree (dorsal), 90 degree (lateral), and 180 degree (ventral) images on standard fluorescent microscopes, such as the Apotome. The technical bottleneck of standard microscopy to scan a whole embryo and larva is also labor-intensive and time-consuming. For instance, fluorescent scanning of a quarter of a 2dpf or 4dpf larvae required about 15 min exposure using a Zeiss Apotome microscope, one imaging modality deployed for this collection. To accelerate the expression profiling of these GBT lines, we subsequently utilized a Zeiss Lightsheet Z.1 SPIM microscope. The Lightsheet enabled high speed scanning of a whole embryo or larva, resulting in a nearly 20x faster image acquisition rate than the primary imaging process utilizing the Apotome. We prioritized and cataloged lines with robust expression at 2 and 4 days post fertilization (dpf). All zebrafish lines are freely available now through zfishbook (Clark et al., 2012) and are partially accessible from the Zebrafish International Resource Center. Consequently, we have generated 1138 lines by using each vector system showing in Supplemental Table 1), and updated results are posted at zfishbook.

Our throughput for cloning the GBT lines using traditional molecular methods was clearly an initial bottleneck. To help address this, we deployed a rapid cloning process based on methods used to isolate retroviral integrations (Varshney, Huang, et al., 2013; Varshney, Lu, et al., 2013). This method leverages the massive parallel sequencing technology of the Illumina MiSeq, yielding 101 bp sequencing reads, followed by a custom bioinformatics pipeline that involves both mapping and annotation. Fin-clips from four male animals per GBT locus are obtained during sperm cryopreservation and are used as a source of DNA. Shared inserts in multiple individuals from a single GBT line are considered candidate loci. This information is subsequently used to generate gene-specific primers to manually confirm linkage using 5‘ or 3’RACE, inverse PCR, or TAIL PCR to generate locus-specific primers for downstream molecular genotyping applications. 212 of the lines met the highest stringency of confirmed expression linkage and are classified as confirmed GBT integrations. An additional 143 lines have been initially tagged using this high throughput method, yielding candidate integration annotation as listed in zfishbook. The ability to complete the annotation status from candidate to confirmed for any given GBT locus with a desired expression profile is enhanced by the continued refinement of the zebrafish genome.

mRFP Expression profiling reveals overlap with known annotation at 2dpf and substantive new expression data at both 2dpf and 4dpf

GBT lines are currently imaged to capture mRFP expression at both 2 and 4 dpf, including dorsal, sagittal and ventral views using the SCORE imaging method. (Petzold et al., 2010) In an openly accessible database of GBT lines, zfishbook, the images of mRFP expression pattern were stored within the media gallery associated with each line (Clark et al., 2012). We summarized published expression data of the tagged genes in both zfishbook and ZFIN in Table 1. Imaging the localization of transcripts

and proteins at 4dpf is more difficult than those at 2 dpf, because accessibility of antisense RNA probes and antibodies into the larva's body is technically limited for in the methods of both *in situ* hybridization and immunohistochemistry. Compared with the published data of gene expression in ZFIN, zfishbook currently provides almost the double number of genes with expression data at 2 dpf and 14 times the number of genes with expression data at 4 dpf (Table 1). In addition, zfishbook also provides novel expression data for 61 genes at any developmental stage (Fig.2).

High knockdown efficiency of endogenous transcripts induced by RP2

We directly compared published transposon insertional mutant vector systems (Fig. 3). The range and average knockdown levels in the FlipTrap system (Trinh le et al., 2011) produced a range of 4-30% (70-96% knockdown capacity) in six tested fish alleles, a similar range to our initial R-series protein trap vectors (R14-R15) that used a simple transcriptional terminator (Liao et al., 2012; Petzold et al., 2009) The pFT1 appears to be an improvement over these systems, in which the overall range and average read-through is reduced to 6-11% (89-94% knockdown) from four tested fish alleles (T. T. Ni et al., 2012). In contrast, the RP2.1 vector (Clark, Balciunas, et al., 2011), (Ding et al., 2013), (Ding et al., 2016) and this manuscript) maintains a strong knockdown (1% or less read-through) in 26 lines tested. Though deployed here using a nearly random, transposon-based delivery platform, the GBT vector system is an effective insertional mutagen suitable for an array of other – including targeted integration - genome-wide applications.

Phenotypic Appearance Rate is Similar to Other Mutagenic Technologies for Forward Genetics Screening through 5 dpf

We conducted an initial forward genetic screen on embryos and early larvae of 179 RFP-positive GBT lines, identifying 12 recessive phenotypes, such as *ryr1b, fras1, tnnt2a, edar and hmcn1*,(Clark, Balciunas, et al., 2011; Westcot et al., 2015) visible during the first five days of development including lethality, heart, muscle, skin and other phenotypes. This 7% recovery of visible early developmental mutants is very similar to the 5% recovered visible mutants from the Sanger TILLING consortium analysis of truncated zebrafish genes (Kettleborough et al., 2013). This 7% is also comparable to prior retroviral (Amsterdam & Hopkins, 2004) and ENU(Haffter et al., 1996) zebrafish mutagenesis work that estimated between 1400 and 2400 genes (~5-9% of the genome) would result in a visible embryonic phenotype when mutated.

GBT alleles phenocopy known embryonic mutations

We tested the first five GBT lines in genes with known loss of function mutant phenotypes (*ryr1b*; *fras1*; *tnnt2a*; *edar; hmcn1*). All five of these alleles in genes with described loss of function defects are phenocopied by these GBT insertional alleles (Clark, Balciunas, et al., 2011); (Westcot et al., 2015); this manuscript). These loci represent a critical internal methods reference further validating the mutagenicity of these novel insertional vectors.

Gene ontology analysis of GBT-tagged loci

To assess the diversity of GBT loci molecularly characterized to date, we utilized the PANTHER classification system (v.14.0, pantherdb.org , (Mi et al., 2013) and generated a table of protein class ontology tags in the molecularly isolated GBT lines. The PANTHER Protein Class ontology was adapted from the PANTHER INDEX (PANTHER/X) ontology that comprises two types of classifications: molecular function and biological process and includes commonly used classes of

protein families. The molecular function schema classifies a protein based on its biochemical properties, such as receptor, cell adhesion molecule, or kinase. The biological process schema classifies a protein based on the cellular role or process in which it is involved, for example, carbohydrate metabolism (cellular role), TCA cycle (pathway), neuronal activities (process), or developmental processes (process) (Thomas PD et al, 2013). As of April 2018, almost half of known zebrafish genes (10626/25289 genes) were tagged in the PANTHER Protein Class ontology. 168 of our cloned GBT alleles mapped in the PANTHER system with 21 types of Protein Classes (Table 2). 18% and 16 % of the mapped GBT alleles are classified to nucleic acid binding (PC00171) and transcription factor (PC00218), respectively (Fig. 4). This result reveals that a quarter of the mapped genes possibly play a role in regulatory processes. Overall, however, the rich diversity of protein classes observed in our cloned traps suggests a large diversity will be represented by the overall collection and consistent with the random nature of genome integration events by the Tol2 transposon (Clark, Balciunas, et al., 2011).

Disease-causing human orthologues of GBT-tagged genes

Of 183 GBT-tagged genes, 171 human orthologues were annotated in at least one public database such as ZFIN, Ensembl, Homologene, and InParanoid (Table 3). Several human orthologues were provisionally annotated using BLASTP and a synteny analysis tool, SynFind. In a previous study comparing the list of human genes possessing at least one zebrafish orthologue with the 3,176 genes bearing morbidity descriptions that are listed in the OMIM database, 82 % morbid genes 2,601 genes) can be related to at least one zebrafish orthologue (Howe et al., 2013). Surprisingly, 67 genes (about 37%) of 183 annotated human orthologues are associated with human disease involved in multi-organ system including nervous, circulatory, endocrine, metabolic, digestive, musculoskeletal, immune, and integument systems (Fig. 5 and Table 3) and many are not established in rodents and zebrafish (Table

5). The GBT protein-trap system provides a variety of potential human disease models which have a revertible allele that can interchange between disease and healthy cellular, organ and physiological states.

The GBT Protein Trap Reveals Protein-Coding Regions Not Predicted by the Zebrafish Genome Project

16 of 211 molecularly confirmed lines by manual, PCR-based mRFP linkage analysis using TALE, inverse, and RACE PCRs (Table 4) do not match any expressed sequence tag (EST) or predicted genes. However, in each case we were able to confirm transcription at the locus in wild-type animals, yielding new annotation for these loci in the zebrafish genome.

Discussions

Novel Expression Annotation Revealed By Protein Trapping of Endogenous Genes

The RP2 and RP8 vectors of GBT system were assembled to capture all three reading frames of fusion protein of the trapped gene with the mRFP reporter. This GBT system reveals that mRFP-truncated fusion proteins exhibit distinct subcellular localization. This is particularly noteworthy for the 4dpf stages because published gene expression data at late developmental stages have been limited by the technical difficulties in conducting such analyses using traditional techniques such as whole mount *in situ* hybridization (WISH) at these larval and later time points. We note that the mRFP-truncated fusion protein may localize ectopically in cases where the protein localization signal is contained in the C-terminal domain (Clark, Balciunas, et al., 2011; Trinh le & Fraser, 2013). Although the extent that subcellular localization recapitulates the endogenous protein is dependent on each insertion locus specifics, the visualizing and illuminating spatiotemporal expression patterns of trapped protein may facilitate dynamic studies of specific cell types and molecular functions in a living vertebrate. The novel expression description at 4dpf in nearly five of six cloned loci demonstrates the dearth of 3D expression annotation for the overwhelming number of genes, even in one of the most studied model system such as the zebrafish. With ever-improving microscope-based imaging tools (Liu et al., 2018), these lines have the potential to help annotate at diverse developmental and adult stages, while also potentially imaging subcellular expression for a subset of protein trap fusions.

Gene-Break Protein Trap in an Effective Insertional Mutagen with High Knockdown Efficiency and Cre-Reversion to WT allele

Transposons offer several unique features over and above traditional static mutational approaches, including high quality expression tools and new regulated mutagenesis methodologies. From the perspective of genome engineering development, GBT technology was the first method for revertible allele generation of vertebrates outside of the mouse (Clark, Balciunas, et al., 2011); (Ding et al., 2013). We know two major potential biases that may yield non-random trapping coverage of the genome. First, the RP2.1 protein trap was initially designed around a single reading frame. Upon molecular analysis of our first lines, however, we discovered that RP2.1 encodes a second, alternative splice acceptor yielding protein trap expression from a second reading frame due to this alternative splicing event in a significant number of our lines. The deliberate development of RP2 and RP8 vectors for each reading frame obviates this potential limitation when used in other delivery modalities besides the Tol2 transposon. For example, these vectors would be suitable for gene editing-based targeted knockin methods.

The GBT system deployed here has been joined by two new and complementary transposon mutagenesis systems. The FlipTrap system by Dr. Fraser's group(Trinh le et al., 2011), which can be mutagenic when provided Cre recombinase, is primarily focused on imaging fusion proteins in vivo and addressing cellular dynamics and related questions. The FT1 system by the Chen lab is a complementary flipping trap that can use either Cre or Flp recombinases to regulate alleles depend on the original orientation of the insertion (T. T. Ni et al., 2012). GBT-based zebrafish alleles are highly complementary and non-redundant to those generated by other mutagenesis methods, including these other transposons and demonstrated higher knockdown efficiency of the WT transcripts than those mutagens because of the use of an enhanced polyadenylation signal and a putative boundary element between 5' protein trap and 3' exon trap cassettes. Importantly, since the initiation of this project to generate a collection of Cre-revertible mutant alleles, several groups have now reported collections of tissue-specific Cre driver lines including the Brand lab (Jungke et al., 2013; http://crezoo.crt-

dresden.de/crezoo/), the Zcre consortium (http://zcre.org.uk/) and the Wen lab in the PTC Consortium . The GBT system described here is a two-component, molecularly regulatable mutagenesis approach that offers the ability to test for the sufficiency of protein-encoding loci in regulated, tissue- and cell- specific applications.

Functional Diversity of the Trapped Proteins by the GBT system

To analyze distribution of protein functions of the trapped genes by GBT system, we performed GO analysis using the PANTHER protein classification. Although the protein functions related in transcriptional regulatory process, such as nucleic acid binding and transcription factors represented one relatively common class of isolated genes, the protein GO analysis indicated that the GBT protein trap was a useful tool for capturing a wide range of protein functions in addition to cell fate regulators and related nuclear genes.

Cloned GBT Loci Represent a Rich Collection of Potential Human Disease Models

More than 7,000 human diseases have already been described, and 80% of those are thought to have a genetic origin (Varga et al., 2018). Model organism studies can be a pivotal resource for understanding gene function which possibly provides additional insight into the cause of particular disease, thereby contributing to understanding of the pathogenic process and discovery of the therapeutic strategy (Wangler et al., 2017). Annotation of human orthologues of 160 tagged genes revealed that GBT technology yield a high frequency potential human disease models with loss of gene function, tracking expression of the truncated protein and Cre-revertible mutated allele to rescue the phenotype.

Furthermore, the modeling human genetic diseases using the GBT system has advantages compared with reverse genetic approaches such as TALEN and CRISPR-Cas9 systems. In initial phenotype screening, PR2.1 mutagenesis demonstrated 7% phenotype appearance as much as the other ENU- and retroviral forward genetic screenings (Amsterdam & Hopkins, 2004; Haffter et al., 1996; Kettleborough et al., 2013). For example, the annotation of disease-causing human orthologues of the tagged genes also revealed that the GBT system comprehensively developed mutants in zebrafish orthologues of human disease loci, including nervous, cardiovascular, endocrine, digestive, musculoskeletal, immune, and integument systems. Surprisingly, this system generated 68 of pioneering mutants in orthologs of human disease loci (Supplemental Table 2).

Discovery of Novel Transcripts by Trapping Unpredicted Genes

Since the completion of the zebrafish reference genome sequencing, it has enabled many new discoveries to be made, in particular the positional cloning of hundreds genes from mutation affecting embryogenesis behavior, physiology, and health and disease. However, a few poorly assembled regions remain (Howe et al., 2013). In molecular cloning of GBT lines generated, we found that a surprising proportion of the sequenced insertions does not correspond to any predicted genes. Although we have not formally excluded that mRFP expression might, in some case, be an artifact, the data of gene prediction provided in genome databases reveals some prediction errors. These results suggest that the algorithms used to predict genes from genome databases have missed a significant number of genes. The protein trapping by using GBT system may useful in identifying unsuspected novel genes, expressions and functions *in vivo* in real time.

New Genomic Insights Using the GBT Random Insertion Mutagen

The ready ease of these mRFP-based protein traps for basic expression analyses demonstrates how much we do not know about our overall proteome and the codex that is our genome. Nearly forty percent and five of six of the cloned genes show no expression at 2dpf and 4dpf, respectively, in ZFIN. At the subcellular level, such protein traps in conjunction with new microscopy techniques represent just another method for cellular and mechanistic analyses in an *in vivo* cellular context. Although these were made using random insertional approaches, new targeted integration tools using gene editing such as GeneWeld(Wierson et al., 2018) should readily empower labs to build their custom GBT lines in the future for genes not in this collection. Together, this initial 1100+ GBT collection is a new contribution to the use of the zebrafish to annotate the vertebrate genome.

Competing interests

The authors declare no competing interests.

Figure legends

Table 1 Comparison availability of expression data

Table 2 Protein classification

Table 3 Disease-causing human orthologues

Table 4 Potential novel human disease models

Table 5 Novel transcripts

Figure 1 Schematic of the RP2 and RP8 gene-break transposon system with all three reading frames of AUG-less mRFP reporter

A. Schematic of the RP2 system fused with 3 reading frames of AUG-less mRFP reporter (RP2.1, RP2.2 and RP2.3). B. Schematic of the RP2 system fused with 3 reading frames of AUG-less mRFP reporter (RP8.1, RP8.2 and RP8.3). ITR, inverted terminal repeat; SA, loxP; Cre recombinase recognition sequence, splice acceptor; *mRFP' AUG-less mRFP sequence; poly (A)+, polyadenylation signal; red octagon, extra transcriptional terminator and putative border element; *β-act*, carp beta-actin enhancer, SD, splice donor; E, enhancer; P, promoter; and WT, wild type.

Figure 2 Novel protein expression

Representative novel expression data of trapped proteins fused to the mRFP reporter in this GBT collection. A. unkl localized in olfactory pit, cerebrum and spinal cord at 4dpf. B. nusap1 localized in retina and the top layer of both forebrain and midbrain at 4 dpf. C. zgc:194659 strongly expressed in the brain and spinal cord at 4 dpf. D. marcksl1a expressed in the lens, skin and notochord at 2 dpf. E. pipp2a specifically localized to otoliths at 2 dpf. F. ahnak specifically expressed in skin at 4 dpf. G. dph1 ubiquitously expressed showing granulized pattern in somites at 4 dpf. H. nfatc3a expressed in skeletal muscle and skin at 4 dpf . I. pard3bb localized to the pronephros and gut at 4 dpf. White arrowhead

shows an artificial expression in lens driven by the promoter of 3' exon trap of RP8.1. J.

LOC100537272 expressed in circulatory cells in the blood stream at 4 dpf.

Figure 3 High knockdown efficiency of RP2.1 compared with other previous gene-trap systems

Black dots of bar graphs shows percentage of remaining endogenous transcripts in homozygous larvae with mean and 95% confidence interval indicated by individual lines. The data of previous protein trap systems were also converted from the data in the original articles, R14-R15, our initial R-series protein trap vectors (n= 6),(Clark, Balciunas, et al., 2011) ; FlipTrap, FlipTrap vectors (n= 6), (Trinh le et al., 2011); FT1, FT1 vector (n=4),(T. T. Ni et al., 2012); RP2.1 (n=26), (Clark, Balciunas, et al., 2011; Ding et al., 2013; El-Rass et al., 2017; Westcot et al., 2015) and unpublished data),

Figure 4 Summary of protein classes categorized the trapped proteins using PANTHER algorism

77 trapped genes were successfully categorized at least one of 21 protein classes by using PANTHER gene ontology algorism. The details of the trapped genes classified in each protein class are listed in Table 2.

Figure 5 Disease-causing human orthologues of the trapped genes involved in human genetic disorders in multi-organ systems.

Supplemental Table 1 Generation of 11 hundred independent lines by all three reading frames of mRFP reporter in both RP2 and RP8 cassettes.

Supplemental Table 2 Disease-causing human orthologues of the trapped protein

Supplemental Figure 1 Representative expression patterns of mRFP fusion protein by RP2 and RP8 integration

Acknowledgements

This work is supported by grants from the National Institutes of Health (GM63904; DA14546; HG006431) and the Mayo Foundation. We thank Zoltan Varga for sharing of the ZIRC sperm cryopreservation protocol prior to publication. Appreciation is also extended to the Mayo Clinic Zebrafish Facility staff for their excellent support. This research was in part funded by the Intramural Research Program of the National Human Genome Research Institute; National Institutes of Health (S.M.B.: 1ZIAHG000183)

References

Amsterdam, A., & Hopkins, N. (2004). Retroviral-mediated insertional mutagenesis in zebrafish. *Methods Cell Biol, 77*, 3-20.

Balciunas, D., Wangensteen, K. J., Wilber, A., Bell, J., Geurts, A., Sivasubbu, S., . . . Ekker, S. C. (2006). Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. *PLoS Genet, 2*(11), e169. doi:10.1371/journal.pgen.0020169

Clark, K. J., Argue, D. P., Petzold, A. M., & Ekker, S. C. (2012). zfishbook: connecting you to a world of zebrafish revertible mutants. *Nucleic Acids Res, 40*(Database issue), D907-911. doi:10.1093/nar/gkr957

Clark, K. J., Balciunas, D., Pogoda, H. M., Ding, Y., Westcot, S. E., Bedell, V. M., . . . Ekker, S. C. (2011). In vivo protein trapping produces a functional expression codex of the vertebrate proteome. *Nat Methods, 8*(6), 506-515. doi:10.1038/nmeth.1606

Clark, K. J., Urban, M. D., Skuster, K. J., & Ekker, S. C. (2011). Transgenic zebrafish using transposable elements. *Methods Cell Biol, 104*, 137-149. doi:10.1016/B978-0-12-374814-0.00008-2

Ding, Y., Liu, W., Deng, Y., Jomok, B., Yang, J., Huang, W., . . . Xu, X. (2013). Trapping cardiac recessive mutants via expression-based insertional mutagenesis screening. *Circ Res, 112*(4), 606-617. doi:10.1161/CIRCRESAHA.112.300603

Ding, Y., Long, P. A., Bos, J. M., Shih, Y. H., Ma, X., Sundsbak, R. S., . . . Xu, X. (2016). A modifier screen identifies DNAJB6 as a cardiomyopathy susceptibility gene. *JCI Insight, 1*(14). doi:10.1172/jci.insight.88797

Draper, B. W., & Moens, C. B. (2009). A high-throughput method for zebrafish sperm cryopreservation and in vitro fertilization. *J Vis Exp*(29). doi:10.3791/1395

El-Rass, S., Eisa-Beygi, S., Khong, E., Brand-Arzamendi, K., Mauro, A., Zhang, H., . . . Wen, X. Y. (2017). Disruption of pdgfra alters endocardial and myocardial fusion during zebrafish cardiac assembly. *Biol Open, 6*(3), 348-357. doi:10.1242/bio.021212

Haffter, P., Granato, M., Brand, M., Mullins, M. C., Hammerschmidt, M., Kane, D. A., . . . Nusslein-Volhard, C. (1996). The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development, 123*, 1-36.

Harris, M. P., Rohner, N., Schwarz, H., Perathoner, S., Konstantinidis, P., & Nusslein-Volhard, C. (2008). Zebrafish eda and edar mutants reveal conserved and ancestral roles of ectodysplasin signaling in vertebrates. *PLoS Genet, 4*(10), e1000206. doi:10.1371/journal.pgen.1000206

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., . . . Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature, 496*(7446), 498-503. doi:10.1038/nature12111

Kawakami, K., Takeda, H., Kawakami, N., Kobayashi, M., Matsuda, N., & Mishina, M. (2004). A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev Cell, 7*(1), 133-144. doi:10.1016/j.devcel.2004.06.005

Kettleborough, R. N., Busch-Nentwich, E. M., Harvey, S. A., Dooley, C. M., de Bruijn, E., van Eeden, F., . . . Stemple, D. L. (2013). A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature, 496*(7446), 494-497. doi:10.1038/nature11992

Liao, H. K., Wang, Y., Noack Watt, K. E., Wen, Q., Breitbach, J., Kemmet, C. K., . . . McGrail, M. (2012). Tol2 gene trap integrations in the zebrafish amyloid precursor protein genes appa and aplp2 reveal accumulation of secreted APP at the embryonic veins. *Dev Dyn, 241*(2), 415-425. doi:10.1002/dvdy.23725

Liu, T. L., Upadhyayula, S., Milkie, D. E., Singh, V., Wang, K., Swinburne, I. A., . . . Betzig, E. (2018). Observing the cell in its native state: Imaging subcellular dynamics in multicellular organisms. *Science, 360*(6386). doi:10.1126/science.aaq1392

Matthews, J. L., Murphy, J. M., Carmichael, C., Yang, H., Tiersch, T., Westerfield, M., & Varga, Z. M. (2018). Changes to Extender, Cryoprotective Medium, and In Vitro Fertilization Improve Zebrafish Sperm Cryopreservation. *Zebrafish, 15*(3), 279-290. doi:10.1089/zeb.2017.1521

Meadows, J. R. S., & Lindblad-Toh, K. (2017). Dissecting evolution and disease using comparative vertebrate genomics. *Nat Rev Genet, 18*(10), 624-636. doi:10.1038/nrg.2017.51

Mi, H., Muruganujan, A., & Thomas, P. D. (2013). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res, 41*(Database issue), D377-386. doi:10.1093/nar/gks1118

Ni, J., Wangensteen, K. J., Nelsen, D., Balciunas, D., Skuster, K. J., Urban, M. D., & Ekker, S. C. (2016). Active recombinant Tol2 transposase for gene transfer and gene discovery applications. *Mob DNA, 7*, 6. doi:10.1186/s13100-016-0062-z

Ni, T. T., Lu, J., Zhu, M., Maddison, L. A., Boyd, K. L., Huskey, L., . . . Chen, W. (2012). Conditional control of gene function by an invertible gene trap in zebrafish. *Proc Natl Acad Sci U S A, 109*(38), 15389-15394. doi:10.1073/pnas.1206131109

Petzold, A. M., Balciunas, D., Sivasubbu, S., Clark, K. J., Bedell, V. M., Westcot, S. E., . . . Ekker, S. C. (2009). Nicotine response genetics in the zebrafish. *Proc Natl Acad Sci U S A, 106*(44), 18662-18667. doi:10.1073/pnas.0908247106

Petzold, A. M., Bedell, V. M., Boczek, N. J., Essner, J. J., Balciunas, D., Clark, K. J., & Ekker, S. C. (2010). SCORE imaging: specimen in a corrected optical rotational enclosure. *Zebrafish, 7*(2), 149-154. doi:10.1089/zeb.2010.0660

Sivasubbu, S., Balciunas, D., Davidson, A. E., Pickart, M. A., Hermanson, S. B., Wangensteen, K. J., . . . Ekker, S. C. (2006). Gene-breaking transposon mutagenesis reveals an essential role for histone H2afza in zebrafish larval development. *Mech Dev, 123*(7), 513-529. doi:10.1016/j.mod.2006.06.002

Stoeger, T., Gerlach, M., Morimoto, R. I., & Nunes Amaral, L. A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol, 16*(9), e2006643. doi:10.1371/journal.pbio.2006643

Trinh le, A., & Fraser, S. E. (2013). Enhancer and gene traps for molecular imaging and genetic analysis in zebrafish. *Dev Growth Differ, 55*(4), 434-445. doi:10.1111/dgd.12055

Trinh le, A., Hochgreb, T., Graham, M., Wu, D., Ruf-Zamojski, F., Jayasena, C. S., . . . Fraser, S. E. (2011). A versatile gene trap to visualize and interrogate the function of the vertebrate proteome. *Genes Dev, 25*(21), 2306-2320. doi:10.1101/gad.174037.111

Urasaki, A., Morvan, G., & Kawakami, K. (2006). Functional dissection of the Tol2 transposable element identified the minimal cis-sequence and a highly repetitive sequence in the subterminal region essential for transposition. *Genetics, 174*(2), 639-649. doi:10.1534/genetics.106.060244

Van Slyke, C. E., Bradford, Y. M., Howe, D. G., Fashena, D. S., Ramachandran, S., Ruzicka, L., & Staff*, Z. (2018). Using ZFIN: Data Types, Organization, and Retrieval. *Methods Mol Biol, 1757*, 307-347. doi:10.1007/978-1-4939-7737-6_11

Varga, M., Ralbovszki, D., Balogh, E., Hamar, R., Keszthelyi, M., & Tory, K. (2018). Zebrafish Models of Rare Hereditary Pediatric Diseases. *Diseases, 6*(2). doi:10.3390/diseases6020043

Varshney, G. K., Huang, H., Zhang, S., Lu, J., Gildea, D. E., Yang, Z., . . . Burgess, S. M. (2013). The Zebrafish Insertion Collection (ZInC): a web based, searchable collection of zebrafish mutations generated by DNA insertion. *Nucleic Acids Res, 41*(Database issue), D861-864. doi:10.1093/nar/gks946

Varshney, G. K., Lu, J., Gildea, D. E., Huang, H., Pei, W., Yang, Z., . . . Lin, S. (2013). A large-scale zebrafish gene knockout resource for the genome-wide study of gene function. *Genome Res, 23*(4), 727-735. doi:10.1101/gr.151464.112

Wangler, M. F., Yamamoto, S., Chao, H. T., Posey, J. E., Westerfield, M., Postlethwait, J., . . . Bellen, H. J. (2017). Model Organisms Facilitate Rare Disease Diagnosis and Therapeutic Research. *Genetics, 207*(1), 9-27. doi:10.1534/genetics.117.203067

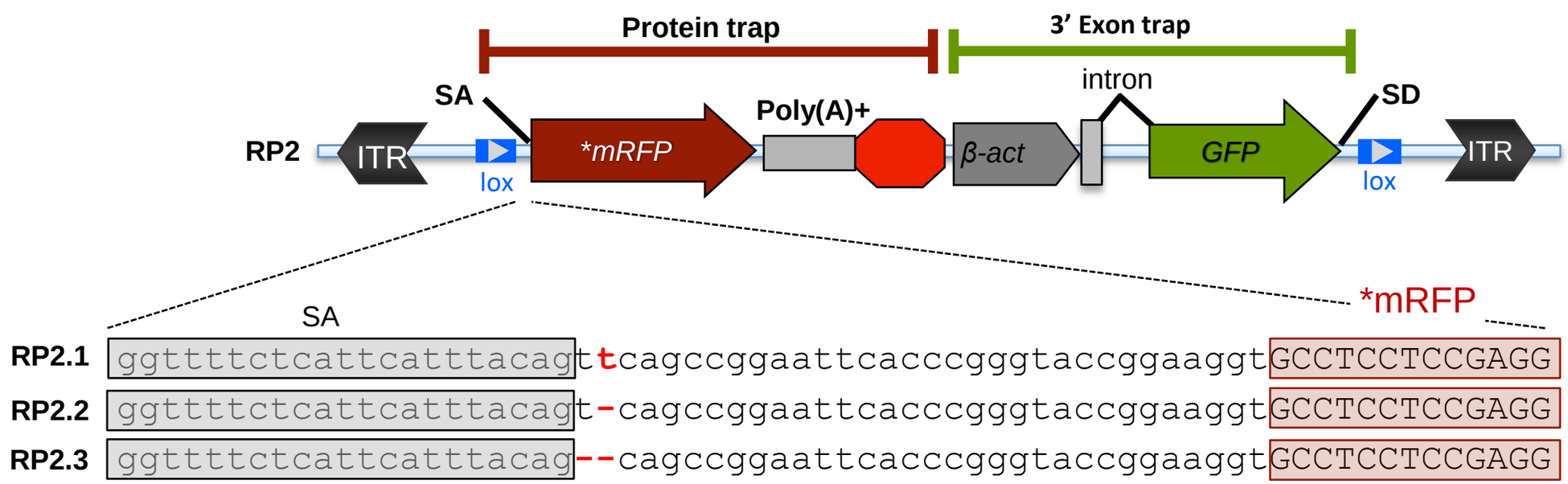
Westcot, S. E., Hatzold, J., Urban, M. D., Richetti, S. K., Skuster, K. J., Harm, R. M., . . . Ekker, S. C. (2015). Protein-Trap Insertional Mutagenesis Uncovers New Genes Involved in Zebrafish Skin Development, Including a

Neuregulin 2a-Based ErbB Signaling Pathway Required during Median Fin Fold Morphogenesis. *PLoS One, 10*(6), e0130688. doi:10.1371/journal.pone.0130688

Wierson, W. A., Welker, J. M., Almeida, M. P., Mann, C. M., Webster, D. A., Torrie, M. E., . . . Essner, J. J. (2018). GeneWeld: a method for efficient targeted integration directed by short homology. *bioRxiv*, 431627. doi:10.1101/431627

Xu, J., Gao, J., Li, J., Xue, L., Clark, K. J., Ekker, S. C., & Du, S. J. (2012). Functional analysis of slow myosin heavy chain 1 and myomesin-3 in sarcomere organization in zebrafish embryonic slow muscles. *J Genet Genomics, 39*(2), 69-80. doi:10.1016/j.jgg.2012.01.005

A



B

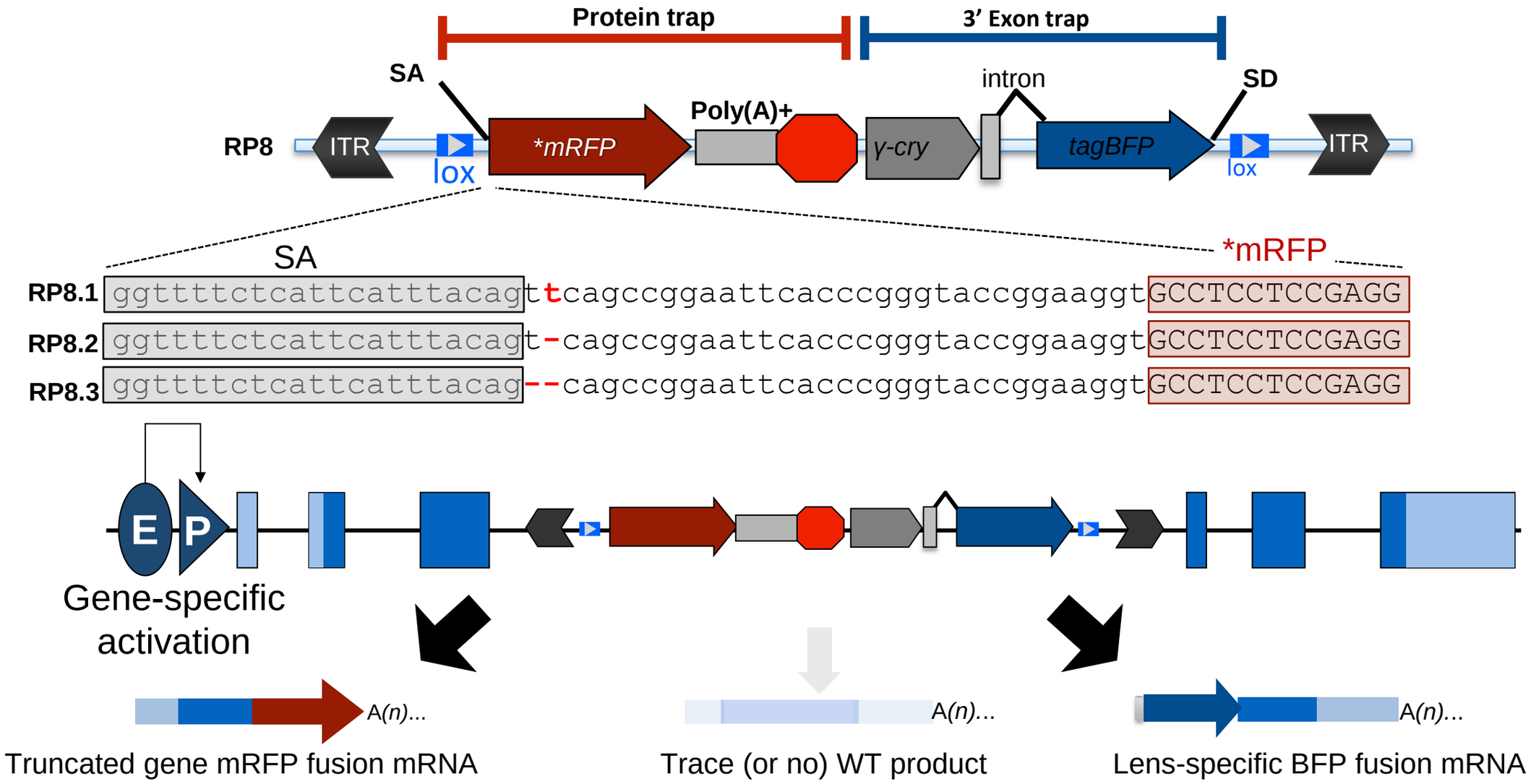


Fig. 1

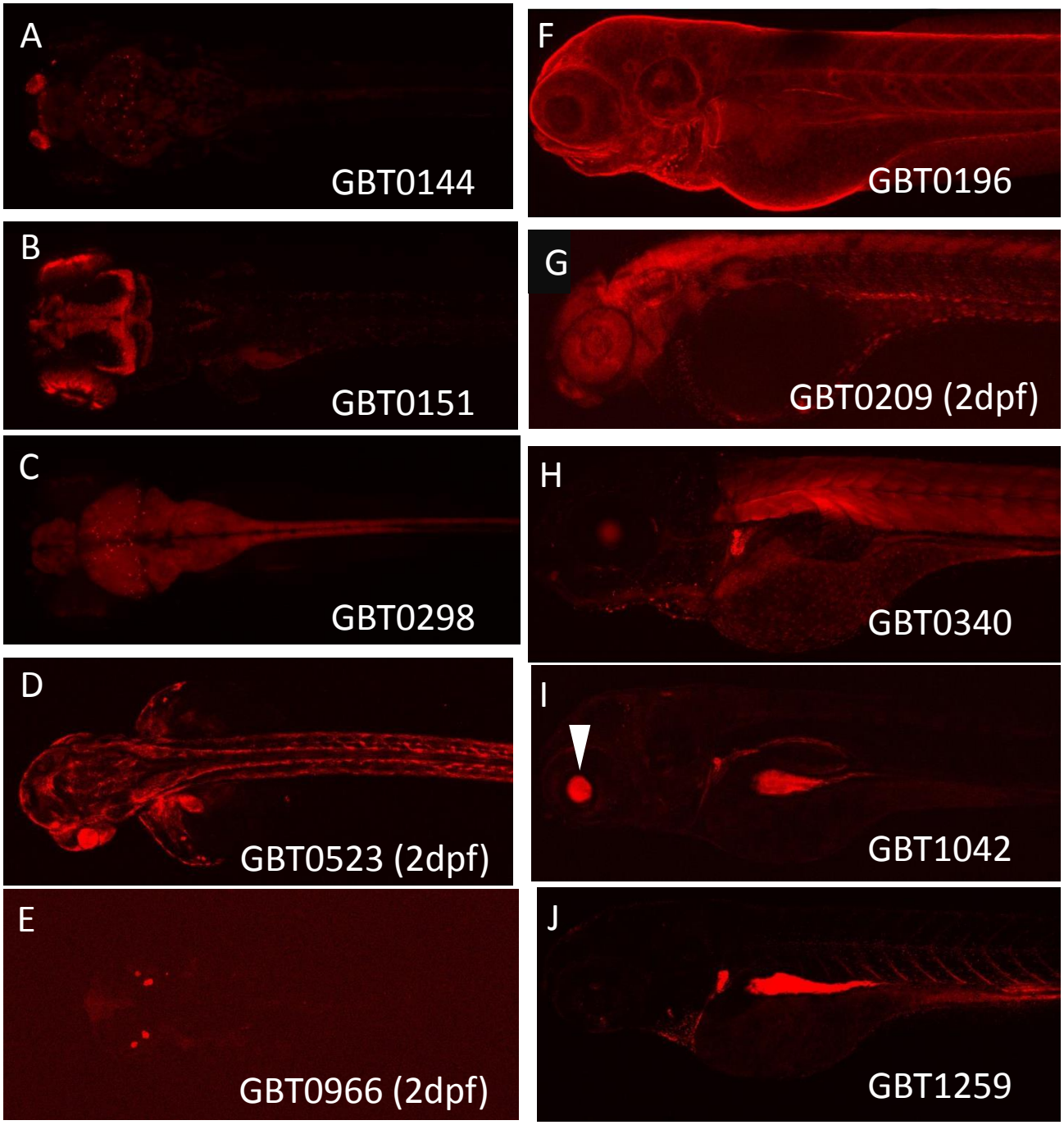

A
GBT0144
F
GBT0196
B
GBT0151
G
GBT0209 (2dpf)
C
GBT0298
H
GBT0340
D
GBT0523 (2dpf)
I
GBT1042
E
GBT0966 (2dpf)
J
GBT1259

Fig.2

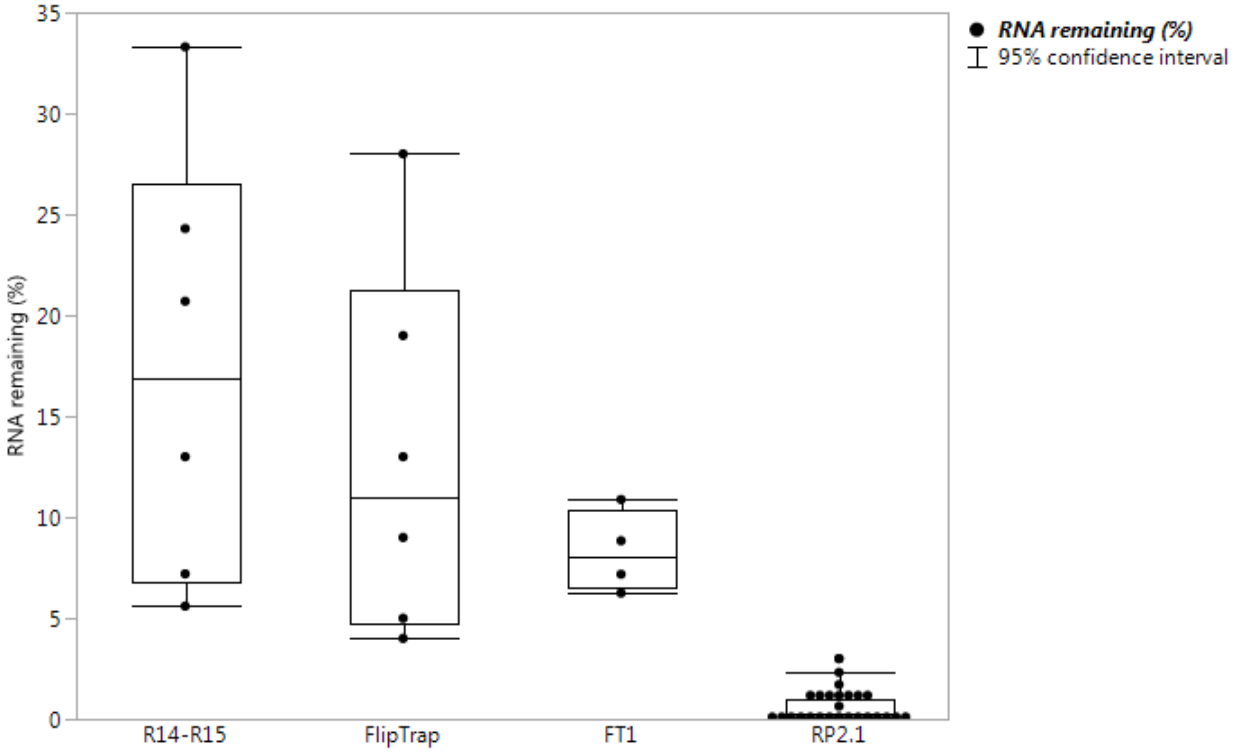

RNA remaining (%)
35
30
25
20
15
10
5
0
R14-R15
FlipTrap
FT1
RP2.1
RNA remaining (%)
95% confidence interval

Fig. 3

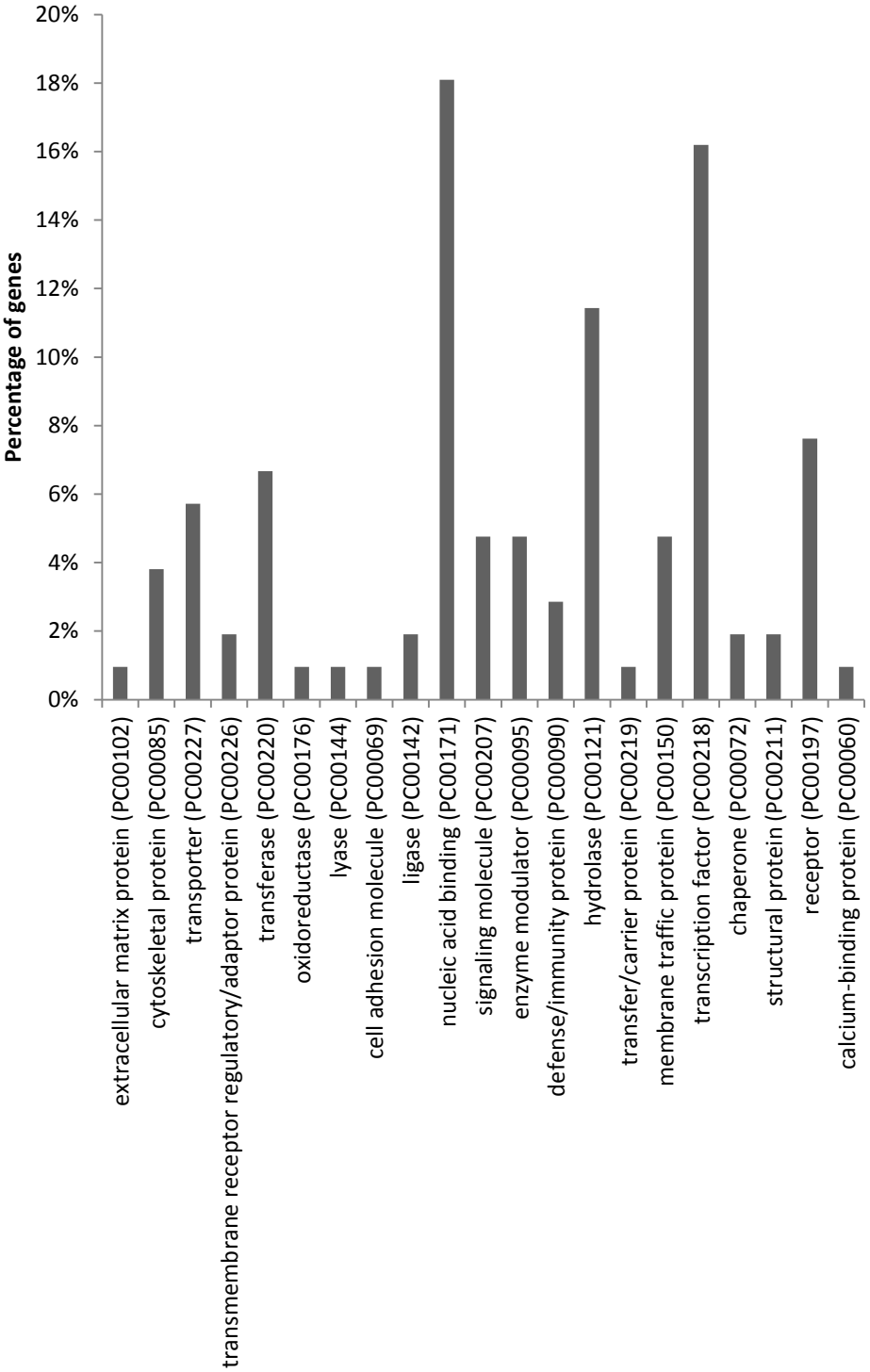
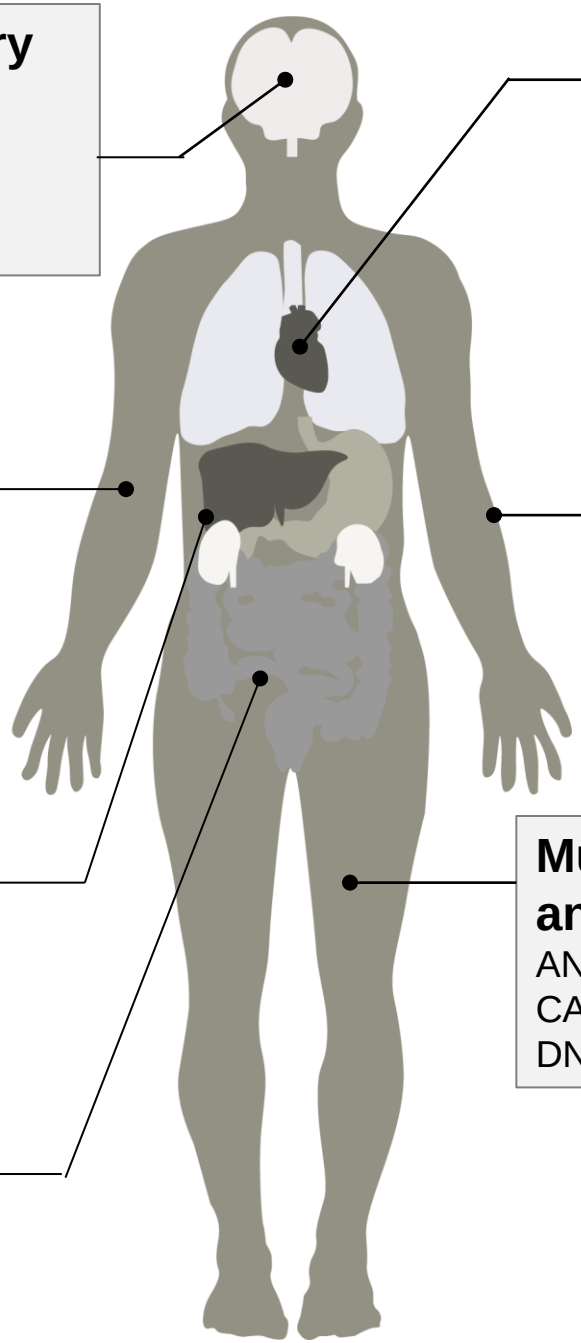

Percentage of genes
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
extracellular matrix protein (PC00102)
cytoskeletal protein (PC00085)
transporter (PC00227)
transmembrane receptor regulatory/adaptor protein (PC00226)
transferase (PC00220)
oxidoreductase (PC00176)
lyase (PC00144)
cell adhesion molecule (PC00069)
ligase (PC00142)
nucleic acid binding (PC00171)
signaling molecule (PC00207)
enzyme modulator (PC00095)
defense/immunity protein (PC00090)
hydrolase (PC00121)
transfer/carrier protein (PC00219)
membrane traffic protein (PC00150)
transcription factor (PC00218)
chaperone (PC00072)
structural protein (PC00211)
receptor (PC00197)
calcium-binding protein (PC00060)

Figure 4



Central, retinal, and auditory nervous system

LRPPRC
EPHA4
PPP3CA
CSNK1D
DLG1
RERE
TENM3
EYA4

Circulatory system

ACTN2
CACNA1C
EYA4
RYR2
TNNT2
TRDN
FLNA
TGFBR1

Immune system

ADA
B2M
RREB1
STAT1

Integument systems

LRMDA
OCA2
KIT
DDB2
EDAR
GRIP1
PTPRF

Endocrine and metabolic system

CACNA1H
GLI2
INSR
MCM9
OGDH

Musculoskeletal system and connective tissue

ANO5
CASQ1
DNAJB6
FKBP10
RYR1
NRP2
TBX15

Digestive system

GPC1

Pleiotropic system

COL4A4
FLNA
FOXL2
FRAS1
GLI2
LRP2
ZFPM2
ZIC2

Fig. 5

Expression data exist	The number of genes	
	2 dpf	4 dpf
zfishbook and ZFIN	45	10
zfishbook	75	155
ZFIN	20	2
Not available	36	9

Table 1

Protein Class	Mapped ID	Gene Symbol	Gene Name
nucleic acid binding (PC00171)			
	ZDB-GENE-040426-893	gar1	H/ACA ribonucleoprotein complex subunit 1
	ZDB-GENE-060503-160	adarb2	Adenosine deaminase, RNA-specific, B2 (non-functional) (Fragment)
	ZDB-GENE-000405-1	pbx1a	Pbx1a homeodomain protein
	ZDB-GENE-060130-4	zfpm2a	Zinc finger protein, FOG family member 2a
	ZDB-GENE-050417-327	eef1a1b	Elongation factor 1-alpha
	ZDB-GENE-081104-328	enox1	Ecto-NOX disulfide-thiol exchanger 1
	ZDB-GENE-050419-169	ddb2	DNA damage-binding protein 2
	ZDB-GENE-980526-306	msxc	Homeobox protein MSH-C
	ZDB-GENE-020529-1	tbx15	T-box 15
	ZDB-GENE-070314-2	rxraa	Retinoic acid receptor RXR-alpha-A
	ZDB-GENE-030131-6117	dido1	Death inducer-obliterator 1
	ZDB-GENE-050913-153	barhl2	BarH-class homeodomain transcription factor
	ZDB-GENE-040426-1272	ddb1	Damage-specific DNA-binding protein 1
	ZDB-GENE-980526-499	stat1a	Signal transducer and activator of transcription
	ZDB-GENE-050517-31	abcf1	ATP-binding cassette, sub-family F (GCN20), member 1
	ZDB-GENE-080403-11	kat2a	Histone acetyltransferase KAT2A
	ZDB-GENE-030131-4505	dhx37	DEAH (Asp-Glu-Ala-His) box polypeptide 37
	ZDB-GENE-060512-241	foxl2a	Forkhead box L2
	ZDB-GENE-041014-310	mcm9	DNA helicase MCM9
transcription factor (PC00218)			
	ZDB-GENE-050419-261	wtip	Wilms tumor protein 1-interacting protein homolog
	ZDB-GENE-050913-139	znf1015	Zinc finger protein 1015
	ZDB-GENE-000405-1	pbx1a	Pbx1a homeodomain protein
	ZDB-GENE-060130-4	zfpm2a	Zinc finger protein, FOG family member 2a
	ZDB-GENE-000710-4	zic2a	ZIC family member 2 (Odd-paired homolog, Drosophila), A
	ZDB-GENE-061207-62	bcl11ba	B cell CLL/lymphoma 11Ba
	ZDB-GENE-050419-146	bhlhe41	BHLH protein DEC2
	ZDB-GENE-030131-6789	taf6l	TAF6-like RNA polymerase II, p300/CBP-associated factor (PCAF)-associated factor
	ZDB-GENE-980526-306	msxc	Homeobox protein MSH-C
	ZDB-GENE-020529-1	tbx15	T-box 15
	ZDB-GENE-070314-2	rxraa	Retinoic acid receptor RXR-alpha-A
	ZDB-GENE-030131-6117	dido1	Death inducer-obliterator 1
	ZDB-GENE-050913-153	barhl2	BarH-class homeodomain transcription factor
	ZDB-GENE-980526-499	stat1a	Signal transducer and activator of transcription
	ZDB-GENE-060130-108	casz1	Castor zinc finger 1
	ZDB-GENE-060512-241	foxl2a	Forkhead box L2
	ZDB-GENE-041111-41	nfatc3a	Nuclear factor of-activated T cells 3a
hydrolase (PC00121)			
	ZDB-GENE-040611-3	nrp2b	Neuropilin
	ZDB-GENE-060503-160	adarb2	Adenosine deaminase, RNA-specific, B2 (non-functional) (Fragment)
	ZDB-GENE-070112-1812	pnpla7a	Patatin-like phospholipase domain-containing 7a
	ZDB-GENE-050417-327	eef1a1b	Elongation factor 1-alpha
	ZDB-GENE-040718-393	ada	Adenosine deaminase
	ZDB-GENE-140703-2	plpp2a	Phospholipid phosphatase 2a
	ZDB-GENE-080818-1	ca16b	Carbonic anhydrase XVI b
	ZDB-GENE-050517-31	abcf1	ATP-binding cassette, sub-family F (GCN20), member 1
	ZDB-GENE-060503-530	ptprf	Receptor-type tyrosine-protein phosphatase F
	ZDB-GENE-070117-757	parga	Poly(ADP-ribose) glycohydrolase
	ENSDARG00000010758	capn12	Calpain 12
	ZDB-GENE-041014-310	mcm9	DNA helicase MCM9
receptor (PC00197)			
	ZDB-GENE-050419-65	v2rl1	Vomeronasal 2 receptor, l1
	ZDB-GENE-060503-5	gabbr1b	Gamma-aminobutyric acid (GABA) B receptor, 1b
	ZDB-GENE-070314-2	rxraa	Retinoic acid receptor RXR-alpha-A
	ZDB-GENE-030131-2427	col7a1	Collagen, type VII, alpha 1
	ZDB-GENE-030909-10	itgb1b	Integrin beta
	ZDB-GENE-080818-1	ca16b	Carbonic anhydrase XVI b
	ZDB-GENE-091027-1	tgfbr1b	Serine/threonine-protein kinase receptor
	ZDB-GENE-070905-3	crfb4	Cytokine receptor family member b4
transferase (PC00220)			
	ZDB-GENE-030131-2049	mat2aa	S-adenosylmethionine synthase
	ZDB-GENE-040426-1516	mboat7	Lysophospholipid acyltransferase 7
	ZDB-GENE-121129-1	csgalnact1a	Hexosyltransferase
	ZDB-GENE-060616-238	mgat5	Mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase
	ZDB-GENE-080403-11	kat2a	Histone acetyltransferase KAT2A
	ZDB-GENE-030131-825	csnk1da	Casein kinase I isoform delta-A
	ZDB-GENE-091027-1	tgfbr1b	Serine/threonine-protein kinase receptor

transporter (PC00227)			
	ZDB-GENE-130103-1	cacna1ha	Voltage-dependent T-type calcium channel subunit alpha
	ZDB-GENE-050419-65	v2rl1	Vomeronasal 2 receptor, l1
	ZDB-GENE-070705-417	ryr1b	Ryanodine receptor 1b (skeletal)
	ZDB-GENE-001127-2	atp1b2a	Sodium/potassium-transporting ATPase subunit beta
	ZDB-GENE-050517-31	abcf1	ATP-binding cassette, sub-family F (GCN20), member 1
	ZDB-GENE-070718-4	oca2	Oculocutaneous albinism II
enzyme modulator (PC00095)			
	ZDB-GENE-060503-160	adarb2	Adenosine deaminase, RNA-specific, B2 (non-functional) (Fragment)
	ZDB-GENE-050417-327	eef1a1b	Elongation factor 1-alpha
	ZDB-GENE-030131-9805	ppp1r13ba	Protein phosphatase 1, regulatory subunit 13Ba
	ZDB-GENE-110411-39	si:dkey-222l13.1	Si:dkey-222l13.1
	ZDB-GENE-040426-1131	phactr4b	Phosphatase and actin regulator 4B
membrane traffic protein (PC00150)			
	ZDB-GENE-030131-5290	napab	N-ethylmaleimide sensitive fusion protein attachment protein alpha
	ZDB-GENE-050522-134	syt5b	Synaptotagmin Vb
	ZDB-GENE-040718-281	vapal	VAMP (vesicle-associated membrane protein)-associated protein A,-like
	ZDB-GENE-050522-235	sncgb	Synuclein, gamma b (breast cancer-specific protein 1)
	ZDB-GENE-110411-39	si:dkey-222l13.1	Si:dkey-222l13.1
signaling molecule (PC00207)			
	ZDB-GENE-091204-186	sh3kbp1	SH3-domain kinase-binding protein 1
	ZDB-GENE-080225-17	arhgef25b	Rho guanine nucleotide exchange factor (GEF) 25b
	ZDB-GENE-070615-10	nrg2a	Neuregulin 2a
	ZDB-GENE-041114-101	fgf13a	Fibroblast growth factor
	ZDB-GENE-040426-1793	fgf13b	Fibroblast growth factor
cytoskeletal protein (PC00085)			
	ZDB-GENE-070112-1512	wasf3b	WAS protein family, member 3b
	ZDB-GENE-080215-23	map7d1b	MAP7 domain-containing 1b
	ZDB-GENE-060503-22	map7d1a	MAP7 domain-containing 1a
	ZDB-GENE-000626-1	tnnt2a	Troponin T type 2a (cardiac)
defense/immunity protein (PC00090)			
	ZDB-GENE-060503-160	adarb2	Adenosine deaminase, RNA-specific, B2 (non-functional) (Fragment)
	ZDB-GENE-040426-2136	b2ml	Beta-2-microglobulin
	ZDB-GENE-070905-3	crfb4	Cytokine receptor family member b4
chaperone (PC00072)			
	ZDB-GENE-040426-876	cct8	Chaperonin-containing TCP1, subunit 8 (theta)
	ZDB-GENE-050522-235	sncgb	Synuclein, gamma b (breast cancer-specific protein 1)
ligase (PC00142)			
	ZDB-GENE-090313-285	zmiz2	Zinc finger, MIZ-type-containing 2
	ZDB-GENE-070912-399	si:dkey-181m9.8	Si:dkey-181m9.8
structural protein (PC00211)			
	ZDB-GENE-030710-8	gpm6ab	Glycoprotein M6Ab (Fragment)
	ZDB-GENE-030710-9	gpm6ba	DMgamma1
transmembrane receptor regulatory/adaptor protein (PC00226)			
	ZDB-GENE-081030-10	cntn3a.1	Contactin 3a, tandem duplicate 1
	ZDB-GENE-010724-8	dlg1	Disks large homolog 1
cell adhesion molecule (PC00069)			
	ZDB-GENE-030909-10	itgb1b	Integrin beta
extracellular matrix protein (PC00102)			
	ZDB-GENE-101112-3	megf6b	Multiple EGF-like-domains 6b
lyase (PC00144)			
	ZDB-GENE-141030-2	npr2	Guanylate cyclase
oxidoreductase (PC00176)			
	ZDB-GENE-081104-328	enox1	Ecto-NOX disulfide-thiol exchanger 1
transfer/carrier protein (PC00219)			
	ZDB-GENE-060628-2	xpo7	Exportin 7
calcium-binding protein (PC00060)			
	ENSDARG00000010758	capn12	Calpain 12

Table 2

Table 2_Protein Classification.xlsx

Line	ZF NCBI gene ID	ZF gene symbol	Human Orthologue NCBI Gene ID	Human Orthologue Gene symbol
GBT0001	559053	casz1	54897	CASZ1
GBT0002	563408	sorbs2b	8470	SORBS2
GBT0005	570216	itgb1b	3688	ITGB1
GBT0007	794348	nrde2	55051	NRDE2
GBT0010	30461	cdh11	1009	CDH11
GBT0016	58138	pbx1a	5087	PBX1
GBT0019	565629	kcnk10b	54207	KCNK10
GBT0020	569437	cntn3a.1	5067	CNTN3
GBT0021	571891	cntnap5b	129684	CNTNAP5
GBT0025	553277	dido1	11083	DIDO1
GBT0028	101882228	CABZ01057928.1	1729	DIAPH1
GBT0031	58071	tnnt2a	7139	TNNT2
GBT0033	100003333	lrch4	4034	LRCH4
GBT0034	567642	si:dkey-181m9.8		
GBT0035	559134	parga	8505	PARG
GBT0038	541513	srpx	8406	SRPX
GBT0039	559579	gabbr1,2	2550	GABBR1
GBT0040	58049	hoxa3a	3201	HOXA4
GBT0043	323266	cd99l2	83692	CD99L2
GBT0046	64270	epha4b	2043	EPHA4
GBT0060	797527	crfb4	3588	IL10RB
GBT0067	559614	myom3	127294	MYOM3
GBT0070	555610	mcm9	254394	MCM9
GBT0073	406467	abcf1	23	ABCF1
GBT0077	406467	abcf1	23	ABCF1
GBT0078	558006	grip1	23426	GRIP1
GBT0082	394037	cct8	10694	CCT8
GBT0091	100002220	enox1	55068	ENOX1
GBT0094	403003	fgf13b	2258	FGF13
GBT0096	394094	sh3glb2b	56904	SH3GLB2
GBT0101	553277	dido1	11083	DIDO1
GBT0103	100002190	cyth3a	9265	CYTH3
GBT0111	569081	rerea	473	RERE
GBT0113	569437	cntn3a.1	5067	CNTN3
GBT0125	100333685	cd302	9936	CD302
GBT0126	405902	nrp2b	8828	NRP2
GBT0131	100004133	ddb2	1643	DDB2
GBT0133	58077	zic2a	7546	ZIC2
GBT0135	563771	bhlhe41	79365	BHLHE41
GBT0137	100004503	eef1a1b	1915	EEF1A1
GBT0141	368225	gpm6ba	2824	GPM6B
GBT0142	100004850	aatka	9625	AATK
GBT0143	566046	wtip	126374	WTIP
GBT0144	565405	unkl	64718	UNKL

Table 3_Human orthologues of GBT confirmed genes updated Apr8-2019.xlsx

GBT0145	560542	epn2	22905	EPN2
GBT0151	567446	nusap1	51203	NUSAP1
GBT0154	449761	si:ch211-163l21.8	57535	KIAA1324
GBT0156	563428	fras1	80144	FRAS1
GBT0157	724006	mgat5	4249	MGAT5
GBT0166	64269	atp1b2a	482	ATP1B2
GBT0168	492758	fgf13a	2258	FGF13
GBT0170	378997	glcci1a	113263	GLCCI1
GBT0172	337397	tmem30aa	55754	TMEM30A
GBT0175	569561	arhgef25b	115557	ARHGEF25
GBT0178	436919	ada	100	ADA
GBT0181	30155	tenm3	55714	TENM3
GBT0186	170581	cacna1c	775	CACNA1C
GBT0187	100007836	si:dkey-253d23.3		
GBT0189	562459	znf414	84330	ZNF414
GBT0190	100001699	barhl2	343472	BARHL2
GBT0195	562459	znf414	84330	ZNF414
GBT0196	559276	ahnak	79026	AHNAK
GBT0200	541428	znrd1	30834	ZNRD1
GBT0201	556341	mhc1zca	3140	MR1
GBT0202	569081	rerea	473	RERE
GBT0203	678534	ntm	150084	IGSF5
GBT0204	492787	cadm4	199731	CADM4
GBT0205	100151756	fam117ab	81558	FAM117A
GBT0208	641575	actn2b	88	ACTN2
GBT0209	550559	dph1	1801	DPH1
GBT0230	797651	jam3a	83700	JAM3
GBT0231	266983	neo1a	4756	NEO1
GBT0232	553567	syt5b	6861	SYT5
GBT0235	557195	lrpprc	10128	LRPPRC
GBT0237	100006951	nrg2a	9542	NRG2
GBT0238	792928	tgfbr1b	7046	TGFBR1
GBT0239	796270	map7d1b	55700	MAP7D1
GBT0240	100093707	bcl11ba	64919	BCL11B
GBT0241	561900	fip1l1a	81608	FIP1L1
GBT0242	554131	tex261	113419	TEX261
GBT0243	563577	ppp1r13ba	23368	PPP1R13B
GBT0245	30526	msx3		
GBT0249	791742	b2ml	567	B2M
GBT0250	386769	ptprma	5797	PTPRM
GBT0251	692279	foxl2a	668	FOXL2
GBT0255	114441	ncam2	4685	NCAM2
GBT0256	550554	ppp3cb	5530	PPP3CA
GBT0261	799247	capn12	147968	CAPN12
GBT0263	559150	hmcn1	83872	HMCN1
GBT0268	100534737	ano5a	203859	ANO5
GBT0270	568996	zfpm2a	23414	ZFPM2

Table 3_Human orthologues of GBT confirmed genes updated Apr8-2019.xlsx

GBT0271	565172	map7d1a	55700	MAP7D1
GBT0275	554270	col4a4	1286	COL4A4
GBT0281	797715	ogdhb	4967	OGDH
GBT0283	793623	sh3kbp1	30011	SH3KBP1
GBT0286	797252	emid1	129080	EMID1
GBT0292	541428	znrd1	30834	ZNRD1
GBT0298	100007373	zgc:194659		
GBT0312	799867	ptprfb	5792	PTPRF
GBT0313	565460	csmd2	114784	CSMD2
GBT0316	324381	fkbp10b	60681	FKBP10
GBT0319	791987	ino80c	125476	INO80C
GBT0321	791987	ino80c	125476	INO80C
GBT0322	101886266	si:ch211-160j14.3	91522	COL23A1
GBT0323	558149	adarb2	105	ADARB2
GBT0325	557764	megf6b	1953	MEGF6
GBT0329	793671	csgalnact1a	55790	CSGALNACT1
GBT0340	561772	nfatc3a	4775	NFATC3
GBT0346	553730	lrmda	83938	LRMDA
GBT0348	570245	ryr1b	6261	RYR1
GBT0357	569183	ca16b	5793	PTPRG
GBT0364	323329	mat2aa	4144	MAT2A
GBT0365	553679	sncgb	6623	SNCG
GBT0369	407734	gpm6ab	2823	GPM6A
GBT0380	795234	znf1140		
GBT0383	554230	kdr	3791	KDR
GBT0389	767723	mosmob	730094	MOSMO
GBT0396	562557	smg9	56006	SMG9
GBT0397	393643	phactr4b	65979	PHACTR4
GBT0398	791173	pnpla7a	375775	PNPLA7
GBT0399	571887	kirrel3a	84623	KIRREL3
GBT0401	560003	ERC1	23085	ERC1
GBT0402	566484	scaf11	9169	SCAF11
GBT0404	564731	CABZ01045212.1	ENSGGOG00000007420	
GBT0409	100331745	npr2	4882	NPR2
GBT0410	436819	vapal	9218	VAPA
GBT0411	393275	dnajb6b	10049	DNAJB6
GBT0412	407710	xpo7	23039	XPO7
GBT0415	570029	arrdc1b	92714	ARRDC1
GBT0416	322795	csrnp1b	64651	CSRNP1
GBT0419	555578	rxraa	6256	RXRA
GBT0422	245700	insrb	3643	INSR
GBT0424	566039	v2rl1		
GBT0425	550455	mrps18b	28973	MRPS18B
GBT0433	393599	ddb1	1642	DDB1
GBT0434	393950	gar1	54433	GAR1
GBT0435	558326	nrxn2a	9379	NRXN2
GBT0437	445226	casq1a	844	CASQ1

Table 3_Human orthologues of GBT confirmed genes updated Apr8-2019.xlsx

GBT0503	30768	stat1a	6772	STAT1
GBT0505	572369	kirrel3b	84623	KIRREL3
GBT0510	553343	si:ch211-266g18.10	10345	TRDN
GBT0511	567419	oca2	4948	OCA2
GBT0513	65239	map2k6	5608	MAP2K6
GBT0520	553367	gpc1a	2817	GPC1
GBT0522	30708	fabp2	2169	FABP2
GBT0523	406407	marcksl1a	65108	MARCKSL1
GBT0525	100333948	mapk8ip1b	9479	MAPK8IP1
GBT0527	386629	tob1a	10140	TOB1
GBT0528	323269	zbtb16a	7704	ZBTB16
GBT0534	327082	napab	8775	NAPA
GBT0545	563632	wasf3b	10810	WASF3
GBT0552	562615	znf1015	163033	ZNF579
GBT0554	492500	gsto2	119391	GSTO2
GBT0570	30154	gli2a	2736	GLI2
GBT0572	568184	lrp2a	4036	LRP2
GBT0585	30256	kita	3815	KIT
GBT0591	100009635	dhx37	57647	DHX37
GBT0593	393716	skp1	6500	SKP1
GBT0599	768182	mcu	90550	MCU
GBT0700	558875	ttc23	64927	TTC23
GBT0710	564112	magi2a	9863	MAGI2
GBT0717	557363	taf6l	10629	TAF6L
GBT0722	569986	dip2cb	22982	DIP2C
GBT0726	100003611	radx	55086	RADX
GBT0734	100149334	si:dkey-15h8.15		
GBT0750	557081	col7a1	1294	COL7A1
GBT0757	555517	kat2a	2648	KAT2A
GBT0760	110437953	adgrl2b.1	23266	ADGRL2
GBT0776	796370	edar	10913	EDAR
GBT0785	393509	mboat7	79143	MBOAT7
GBT0795	562459	znf414	84330	ZNF414
GBT0906	393638	timm50	92609	TIMM50
GBT0926	790941	dele1	9812	DELE1
GBT0936	541552	eya4	2070	EYA4
GBT0941	100333571	zmiz2	83637	ZMIZ2
GBT0951	393521	macrod2	140733	MACROD2
GBT0959	562529	flna	2316	FLNA
GBT0966	563806	plpp2a	8612	PLPP2
GBT0972	114446	dlg1	1739	DLG1
GBT0978	568996	zfpm2a	23414	ZFPM2
GBT0980	559150	hmcn1	83872	HMCN1
GBT0993	557073	cdip1	29965	CDIP1
GBT1023	322106	csnk1da	1453	CSNK1D
GBT1027	557315	kirrel3l	84063	KIRREL2
GBT1033	555701	rreb1b	6239	RREB1

Table 3_Human orthologues of GBT confirmed genes updated Apr8-2019.xlsx

GBT1042	562146	pard3bb	117583	PARD3B
GBT1093	100126126	ryr2a	6262	RYR2
GBT1105	246222	tbx15	6913	TBX15
GBT1129	560875	cacna1ha	8912	CACNA1H
GBT1248	100148041	tnk2a	10188	TNK2
GBT1259	100537272	LOC100537272		
GBT1278	735249	selenos	55829	SELENOS
GBT1300	386856	pdgfra	5156	PDGFRA

Table 3

Table 3_Human orthologues of GBT confirmed genes updated Apr8-2019.xlsx

GBT	Tagged Gene	Human Orthologue	Human Disease	Disease Models (Number of models)	Reference (PMID)
GBT0016	pbx1a	PBX1	CONGENITAL ANOMALIES OF KIDNEY AND URINARY TRACT SYNDROME WITH ORWITHOUT HEARING LOSS ABNORMAL EARS OR DEVELOPMENTAL DELAY	mouse (1)	12591246
GBT0031	tnnt2a	TNNT2	CARDIOMYOPATHY DILATED 1D	mouse (5)	18606313, 27936050, 17556660, 18349139
GBT0031	tnnt2a	TNNT2	CARDIOMYOPATHY FAMILIAL HYPERTROPHIC 2	mouse (9)	16326803, 9788962, 11171784, 18349139, 10449439, 9637714, 23532597
GBT0078	grip1	GRIP1	FRASER SYNDROME 3	mouse (2)	14730302, 16880404
GBT0131	ddb2	DDB2	XERODERMA PIGMENTOSUM COMPLEMENTATION GROUP E	mouse (4)	14769931, 15558025
GBT0133	zic2a	ZIC2	HOLOPROSENCEPHALY 5	mouse (3)	27466203, 18617531, 10677508
GBT0135	bhlhe41	BHLHE41	SHORT SLEEPER	mouse (1)	19679812
GBT0156	fras1	FRAS1	FRASER SYNDROME 1	mouse (5)	12766769, 12766770, 24143185, 15623520, 15838507
GBT0178	ada	ADA	SEVERE COMBINED IMMUNODEFICIENCY AUTOSOMAL RECESSIVE TCELL-NEGATIVE B CELL-NEGATIVE NK CELL-NEGATIVE DUE TO ADENOSINEDEAMINASE DEFICIENCY	mouse (1)	9478961
GBT0186	cacna1c	CACNA1C	TIMOTHY SYNDROME	mouse (1)	21878566
GBT0240	bcl11ba	BCL11B	IMMUNODEFICIENCY 49	zebrafish (1)	
GBT0251	foxl2a	FOXL2	Blepharophimosis ptosis and epicanthus inversus	mouse (3)	15056605, 14736745, 24565867
GBT0268	ano5a	ANO5	MUSCULAR DYSTROPHY LIMB-GIRDLE AUTOSOMAL RECESSIVE 12	mouse (1)	26911675
GBT0270	zfpm2a	ZFPM2	Tetralogy of Fallot	mouse (1)	10892744, 12223418
GBT0270	zfpm2a	ZFPM2	DIAPHRAGMATIC HERNIA 3	mouse (1)	16103912
GBT0275	col4a4	COL4A4	ALPORT SYNDROME AUTOSOMAL RECESSIVE	mouse (4)	21196518, 24522496
GBT0348	ryr1b	RYR1	CENTRAL CORE DISEASE OF MUSCLE	mouse (3)	25564733, 19959667, 7515481
GBT0396	smg9	SMG9	HEART AND BRAIN MALFORMATION SYNDROME	mouse (1)	27018474
GBT0409	npr2	NPR2	Acromesomelic dysplasia Maroteaux type	mouse (2)	23065701, 17728275
GBT0411	dnajb6b	DNAJB6	autosomal dominant limb-girdle muscular dystrophy type 1	zebrafish (1-NOT), mouse (1)	26362252
GBT0422	insrb	INSR	DONOHUE SYNDROME	mouse (2-NOT)	
GBT0511	oca2	OCA2	ALBINISM OCULOCUTANEOUS TYPE II	rabbit (1)	Magnussen, K. (1952).
GBT0572	lrp2a	LRP2	DONNAI-BARROW SYNDROME	mouse (1)	20653565
GBT0585	kita	KIT	MASTOCYTOSIS CUTANEOUS	mouse (5)	24788138, 21148330
GBT0585	kita	KIT	Piebald trait	mouse (1)	20095975
GBT0585	kita	KIT	Gastrointestinal stromal tumor	mouse (5)	16061643, 22652566, 18098338, 12754375
GBT0710	magi2a	MAGI2	NEPHROTIC SYNDROME TYPE 15	mouse (1)	25271328
GBT0750	col7a1	COL7A1	recessive dystrophic epidermolysis bullosa	mouse (4)	18382769, 19893033, 10523500
GBT0776	edar	EDAR	hypohidrotic ectodermal dysplasia	mouse (1)	10431242, 17148670, 9799835
GBT0776	edar	EDAR	HAIR MORPHOLOGY 1	mouse (1)	18561327
GBT0785	mboat7	MBOAT7	MENTAL RETARDATION AUTOSOMAL RECESSIVE 57	mouse (1)	23097495
GBT0959	flna	FLNA	Periventricular nodular heterotopia 1	mouse (1)	16825286
GBT1023	csnk1da	CSNK1D	ADVANCED SLEEP PHASE SYNDROME FAMILIAL 2	mouse (2)	15800623
GBT1093	ryr2a	RYR2	Ventricular tachycardia catecholaminergic polymorphic 1 with or without atrial dysfunction and/or dilated cardiomyopathy	mouse (10)	27482086, 25775566, 26121139, 24755079, 23152493, 22828895, 20224043, 18419777, 16873551, 15890976
GBT1093	ryr2a	RYR2	ARRHYTHMOGENIC RIGHT VENTRICULAR DYSPLASIA FAMILIAL 2	mouse (1)	16873551
GBT1105	tbx15	TBX15	COUSIN SYNDROME	mouse (1)	19068278

Table 4

Line	Vector	Genome location	Genome version
GBT0115	RP2	chr8:12923125-12923133	ZV9
GBT0129	RP2	chr12:45451589-45451597	ZV9
GBT0148	RP2	chr16:55188302-55188310	ZV9
GBT0264	RP2	chr5:49960612-49960620	ZV9
GBT0506	RP2	chr18:46186350-46186358	ZV9
GBT0573	RP2	chr6:33630487-33630495	GRCz10
GBT0586	RP2	chr13:9112806-9112814	GRCz10
GBT0702	RP2	chr13:9112806-9112814	GRCz10
GBT0724	RP8	chr7:7633071-7633079	GRCz10
GBT0994	RP2	chr3:39627361-39627369	GRCz10
GBT1024	RP2	chr25:5175242-5175250	GRCz10
GBT1071	RP8	chr9:21818261-21818269	GRCz10
GBT1100	RP2	chr5:25509243-25509251	GRCz10
GBT1116	RP2	chr14:14724098-14724106	GRCz10
GBT1168	RP2	chr5:39664405-39664412	GRCz10

Table 5