A defined zebrafish line for high-throughput genetics and genomics: NHGRI-1

Authors: Matthew C. LaFave¹, Gaurav K. Varshney¹, Meghana Vemulapalli², James C. Mullikin²,³, Shawn M. Burgess¹*

¹) Developmental Genomics Section, Translational and Functional Genomics Branch, NHGRI, NIH, Bethesda, MD 20892
²) NIH Intramural Sequencing Center, NHGRI, NIH, Bethesda, MD 20814
³) Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, NHGRI, NIH, Bethesda, MD 20814

* Correspondence: burgess@mail.nih.gov

Keywords: zebrafish; SNV; genome sequence; CRISPR; variants

Abstract

Substantial intra-strain variation at the nucleotide level complicates molecular and genetic studies in zebrafish, such as the use of CRISPRs or morpholinos to inactivate genes. In the absence of robust inbred zebrafish lines, we generated NHGRI-1, a healthy and fecund strain derived from founder parents we sequenced to a depth of ~50x. Within this strain, we have identified the majority of the genome that matches the reference sequence, and documented most of the variants. This strain has utility for many reasons, but in particular it will be useful for any researcher who needs to know the exact sequence (with all variants) of a particular genomic region, or who wants to be able to robustly map sequences back to a genome with all possible variants defined.

The zebrafish (Danio rerio) is a powerful tool for understanding vertebrate biology. The usefulness of this model organism is bolstered by the availability of
a “finished” sequenced and annotated genome (FLICEK et al. 2014; HOWE et al. 2013). As a natural extension of this resource, there are several high-throughput efforts to systematically mutagenize all zebrafish protein-coding genes (KETTLEBOROUGH et al. 2013; MOENS et al. 2008; VARSHNEY et al. 2013a; VARSHNEY et al. 2013b).

In addition to such projects, the combination of a sequenced genome and developments in targeted nuclease technology mean that the zebrafish community is now able to rapidly take advantage of custom genome editing technologies (BEDELL et al. 2012; DOYON et al. 2008; HRUSCHA et al. 2013; HWANG et al. 2013; JAO et al. 2013). CRISPRs in particular provide an efficient, easy, and inexpensive means of manipulating and interrogating the genome (CONG et al. 2013; JINEK et al. 2012; MALI et al. 2013). However, because there are very few hardy inbred zebrafish lines (over-inbreeding tends to result in unhealthy stocks) and polymorphism rates are close to one every one hundred bases, variants frequently have the potential to interfere with target site design (BOWEN et al. 2012; GURYEV et al. 2006; STICKNEY et al. 2002) or with regions of homology used for homologous recombination. In general, genome targeting is heavily dependent on an exact match to the primary sequence. Depending on the sequence, even a single mismatch can severely reduce the cutting efficiency (HSU et al. 2013). In addition, other techniques such as RNA-Seq or ChIP-Seq are substantially less accurate without having fully characterized variants in the background strain. Therefore, it is preferable to carry out studies in a zebrafish
strain in which the regions of invariant sequence are known with a high degree of confidence and all variants are categorized to allow for robust genomic mapping.

With these concerns in mind, we derived the zebrafish line NHGRI-1. NHGRI-1 fish were derived from an original strain known as “TAB-5” made from a hybrid cross between fish from two of the most commonly used zebrafish lines: Tübingen and AB (Haffter et al. 1996; Streisinger et al. 1981). The F1 fish from this cross were inbred and screened to be clear of any mutations impacting the first five days of development. Since its initial isolation in 1997, we have carried the strain in the lab until the present day without introducing other outside genetic diversity. We selected several mating pairs from the TAB-5 pool, and the most robust mating pair was chosen as the founding pair for NHGRI-1. We are now on the 3rd generation of NHGRI-1 and their fecundity and overall health remain strong.

We carried out high-throughput sequencing to a depth of ~50x for each parent. The male and female sequencing libraries had a combined 1,289,142,362 non-duplicate reads, with a median coverage of 52x and 47x, respectively. By doing so, we identified over 10 million previously unreported single nucleotide variants (SNVs). The raw sequence data has been deposited in SRA [BioProject ID: 246102]. In addition, we have identified nearly all the regions of the genome that are invariant relative to the Zv9 reference sequence. We generated a BED file of invariant nucleotides, which indicates the regions in which there was both a lack of alternative alleles and sufficient read depth and genotype confidence to call bases as invariant (Fig. 1). 71% of the genome fits
these criteria. The invariant file is hosted on the NHGRI-1 website at
http://research.nhgri.nih.gov/manuscripts/Burgess/zebrafish/download.shtml, a
UCSC data hub has been established at
http://research.nhgri.nih.gov/manuscripts/Burgess/zebrafish/downloads/NHGRI-
1/hub.txt, and data has been transferred to ZFIN.org. Information on the variants
themselves can be downloaded from dbSNP [submitter handle: NHGRI_DGS;
submitter batch ID: NHGRI-1_founders]. The invariant regions are easily
identified by using the BED file, simplifying the design of CRISPR targets,
amplicon primers, finding regions for homologous recombination, Morpholino
design, or essentially any experiment that requires high confidence in the exact
sequence of the genomic region of interest.

We detected over 17 million total variants upon merging the variant calls
from the two libraries. Of that total, 236,301 were in exons of Ensembl transcripts
(Table 1). Variants were called as homozygous only if they were homozygous in
both fish; such variants will stably retain the variant allele in future generations.

To underscore the issues related to background variation in the commonly
used zebrafish lines, we detected 669 variants that formed premature stop
codons in at least one transcript, 105 of which were homozygous mutant in both
sexes. We have generated a BED track of these variants, indicating the location,
alternative allele, and homo/heterozygosity available from the NHGRI-1 website
(Fig. 1). A list of affected genes can also be found in Table S1.

We detected 3,160 deletion or insertion variants (DIVs) in exons. DIVs of a
length divisible by three were highly represented and comprise about 60% of the
DIVs (Fig. 2A). Presumably, this is because the resultant non-frameshift mutations would be less likely to be selected against than those that produce frameshifts. A similar profile has been reported in human indels (Chen et al. 2007). This trend is not present in the genome-wide set of 2,210,080 NHGRI-1 DIVs (Fig. 2B).

We compared the SNVs identified in NHGRI-1 with dbSNP (Build ID: 139) and a publicly available dataset obtained from low-coverage sequencing of multiple zebrafish lines (Bowen et al. 2012; Sherry et al. 2001). For simplicity, we only compared bi-allelic SNVs for which the reference sequence is known (i.e., no “N”s). The majority of NHGRI-1 SNVs had not been previously reported in either dataset (Fig. 3). We find that the rate of SNVs per sequenced base in NHGRI-1 is 0.01, or about 12.5-20x higher than the rate in humans (Kidd et al. 2008). It is important to note that, while the 0.01 figure is relevant for NHGRI-1, the regions of homozygosity created by inbreeding mean it certainly underestimates the SNV load in zebrafish as a whole.

We also compared the mutational profile of NHGRI-1 to that reported for a zebrafish captured from the wild and sequenced at 39x coverage (Patowary et al. 2013). Different cutoffs had been applied for variant calling in said study, such as a minimum of 32 reads to call an SNV and 5 reads to call a DIV, but the ratios of variant types can be compared. The differences are statistically significant, but small. Among the SNVs in the wild zebrafish, 22.3% were reported as being homozygous, compared to 17.8% in NHGRI-1 (Fisher's exact test, $P < 2.2 \times 10^{-16}$). Deletions are more prevalent than insertions in both studies, with the wild
zebrafish reported as having 53.9% deletions, compared to 51.6% in NHGRI-1 ($P < 2.2 \times 10^{-16}$).

This fish line will have utility in terms of automated design for targeted nucleases, as well as for studies such as ChIP-seq or RNA-seq where SNVs or DIVs might reduce the accuracy of mapping the raw sequence data. In addition, techniques such as homologous recombination are very sensitive to variants (TIELE et al. 1992), NHGRI-1 will allow researchers to target genomic regions that do not contain any variant nucleotides. Thus, NHGRI-1 will prove useful in a variety of circumstances where absolute knowledge of the possible sequence variation is needed. The line will be distributed by the Zebrafish International Resource Center (http://zebrafish.org) and the European Zebrafish Resource Center (http://www.ezrc.kit.edu).

Acknowledgement

This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

Literature Cited

BEDELL, V. M., Y. WANG, J. M. CAMPBELL, T. L. POSHUSTA, C. G. STARKER et al., 2012 In vivo genome editing using a high-efficiency TALEN system. Nature 491: 114-118.
BOWEN, M. E., K. HENKE, K. R. SIEGFRIED, M. L. WARMAN and M. P. HARRIS, 2012 Efficient mapping and cloning of mutations in zebrafish by low-coverage whole-genome sequencing. Genetics 190: 1017-1024.
CHEN, F. C., C. J. CHEN, W. H. LI and T. J. CHUANG, 2007 Human-specific insertions and deletions inferred from mammalian genome sequences. Genome research 17: 16-22.

CONG, L., F. A. RAN, D. COX, S. LIN, R. BARRETT et al., 2013 Multiplex genome engineering using CRISPR/Cas systems. Science 339: 819-823.

DOYON, Y., J. M. MCCAMMON, J. C. MILLER, F. FARAJI, C. NGO et al., 2008 Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. Nature biotechnology 26: 702-708.

FliceK, P., M. R. AMODE, D. BARRELL, K. BEAL, K. BILLIS et al., 2014 Nucleic acids research 42: D749-755.

Guryev, V., M. J. KOUDIJS, E. BEREZIKOV, S. L. JOHNSON, R. H. PLASTERK et al., 2006 Genetic variation in the zebrafish. Genome research 16: 491-497.

Haffter, P., M. GRANATO, M. BRAND, M. C. MULLINS, M. HAMMERSCHMIDT et al., 1996 The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. Development 123: 1-36.

Howe, K., M. D. CLARK, C. F. TORRuja, J. TORMRANCE, C. BERTHELOT et al., 2013 The zebrafish reference genome sequence and its relationship to the human genome. Nature 496: 498-503.

HRUSCHA, A., P. KRAWITZ, A. RECHENBERG, V. HEINRICH, J. HECHT et al., 2013 Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. Development 140: 4982-4987.

HSU, P. D., D. A. SCOTT, J. A. WEINSTEIN, F. A. RAN, S. KONERMANN et al., 2013 DNA targeting specificity of RNA-guided Cas9 nucleases. Nature biotechnology 31: 827-832.

HWANG, W. Y., Y. FU, D. REYON, M. L. MAEDER, S. Q. TSAI et al., 2013 Efficient genome editing in zebrafish using a CRISPR-Cas system. Nature biotechnology 31: 227-229.

JAO, L. E., S. R. WENTE and W. CHEN, 2013 Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. Proceedings of the National Academy of Sciences of the United States of America 110: 13904-13909.

Jinek, M., K. CHYLINSKI, I. FONFARA, M. HAUSER, J. A. DOUDNA et al., 2012 A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337: 816-821.

KETTLEBOROUGH, R. N., E. M. BUSCH-NENTWICH, S. A. HARVEY, C. M. DOOLEY, E. DE BRUIJN et al., 2013 A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature 496: 494-497.

KIDD, J. M., G. M. COOPER, W. F. DONAHUE, H. S. HAYDEN, N. SAMPS et al., 2008 Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56-64.

Li, H., B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUA et al., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

Mali, P., L. YANG, K. M. ESVELT, J. AACH, M. GUELL et al., 2013 RNA-guided human genome engineering via Cas9. Science 339: 823-826.
MOENS, C. B., T. M. DONN, E. R. WOLF-SAXON and T. P. MA, 2008 Reverse genetics in zebrafish by TILLING. Briefings in functional genomics & proteomics 7: 454-459.

PATOWARY, A., R. PURKANTI, M. SINGH, R. CHAUHAN, A. R. SINGH et al., 2013 A sequence-based variation map of zebrafish. Zebrafish 10: 15-20.

SHERRY, S. T., M. H. WARD, M. KHOLODOV, J. BAKER, L. PHAN et al., 2001 dbSNP: the NCBI database of genetic variation. Nucleic acids research 29: 308-311.

STICKNEY, H. L., J. SCHMUTZ, I. G. WOODS, C. C. HOLTZER, M. C. DICKSON et al., 2002 Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays. Genome research 12: 1929-1934.

STREISINGER, G., C. WALKER, N. DOWER, D. KNAUBER and F. SINGER, 1981 Production of clones of homozygous diploid zebra fish (Brachydanio rerio). Nature 291: 293-296.

TE RIELE, H., E. R. MAANDAG and A. BERNs, 1992 Highly efficient gene targeting in embryonic stem cells through homologous recombination with isogenic DNA constructs. Proceedings of the National Academy of Sciences of the United States of America 89: 5128-5132.

TEER, J. K., L. L. BONNYCASTLE, P. S. CHINES, N. F. HANSEN, N. AOYAMA et al., 2010 Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome research 20: 1420-1431.

VARSHNEY, G. K., H. HUANG, S. ZHANG, J. LU, D. E. GILDEA et al., 2013a The Zebrafish Insertion Collection (ZInC): a web based, searchable collection of zebrafish mutations generated by DNA insertion. Nucleic acids research 41: D861-864.

VARSHNEY, G. K., J. LU, D. E. GILDEA, H. HUANG, W. PEI et al., 2013b A large-scale zebrafish gene knockout resource for the genome-wide study of gene function. Genome research 23: 727-735.

WANG, K., M. Li and H. HAKONARSON, 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38: e164.
Figure 1. Screenshot of the UCSC browser custom tracks for NHGRI-1.

Twenty mating pairs from six-month-old TAB-5 fish were screened to select a robust founding pair with good clutch size and healthy progeny; the most fecund pair was renamed NHGRI-1. Fin clips from the NHGRI-1 male and female were prepared as separate genomic DNA libraries and sequenced on the Illumina HiSeq 2000 by the NIH Intramural Sequencing Center. Both libraries were subjected to paired-end sequencing with 101 bp reads. We aligned the sequence to the zebrafish genome (Zv9, Howe et al. 2013) with Novoalign version 2.08.02 (http://www.novocraft.com/). We removed PCR duplicates via SAMtools version 0.1.18 (Li et al. 2009). We used bam2mpg to identify the most probable genotype (MPG) for nucleotides in both parents (Teer et al. 2010). Bases that did not have an MPG score of at least 10, coverage of at least 20x, and a ratio of MPG score to coverage greater than 0.5 were discarded. Regions of low sequence complexity were not specifically excluded from the analysis unless they failed to meet these criteria. The bases that matched the reference and met the above criteria in both fish were used to build the BED track of invariant nucleotides. The top track indicates the bases that were invariant in both fish sequenced. The white regions indicate either variation in at least one fish, or insufficient read depth to confidently call the region as invariant. The second track indicates two nonsense mutations detected in this region. The letter indicates the alternative allele, and the color indicates if the mutation was homozygous (red) or heterozygous (blue) in the NHGRI-1 population. Both tracks are available as a track hub, which is hosted at
Figure 2. Deletion and insertion variant length distribution within exons. A. The 3,160 DIVs in exons. B. The 2,210,080 DIVs detected genome-wide. Red bars indicate the number of deletions of a given length; blue bars represent insertions.

Figure 3. SNV overlap with publically available datasets. This comparison only incorporates SNVs that were bi-allelic, and for which the reference base was an unambiguous A, C, G, or T. The Bowen et al. SNVs were downloaded from http://fishbonelab.org/harris/Resources_files/parental_variants.tar; both datasets were downloaded on March 12th, 2014.
|                  | SNV   | DIV   | Total  |
|------------------|-------|-------|--------|
| Total variants   | 14917339 | 2210080 | 17127419  |
| Heterozygous     | 12245715 | 1953277 | 14198992  |
| Homozygous       | 2642908 | 225347  | 2868255   |
| Unknown          | 28716   | 31456  | 60172    |
| Exon variants    | 233141  | 3160   | 236301   |
| Heterozygous     | 190626  | 2815   | 193441   |
| Homozygous       | 42153   | 311    | 42464    |
| Unknown          | 362     | 34     | 396      |

Table 1. Raw counts of variants in NHGRI-1. Single nucleotide variants and deletion and insertion variants were annotated using ANNOVAR version 2012-10-16 (Wang et al. 2010). Our annotation used the ensGene track hosted on the UCSC genome browser, which corresponded to Ensembl release 74 (Flieck et al. 2014). We annotated the male and female fish separately, then combined the ANNOVAR output to determine overall homozygosity and heterozygosity. Variants were only considered homozygous in NHGRI-1 if they were independently called as homozygous in both sexes. We identified a variant as unknown if it was called as 1) unknown in both sexes, or 2) unknown in one fish and homozygous reference in the other. All remaining variants were considered to be heterozygous in NHGRI-1, even if they were called as homozygous in one of the sexes. In cases in which deletion or insertion variants (DIVs) of different lengths were reported at the same position, both were counted as separate variants.
Table 2. Mutations introduced by variants in NHGRI-1.

| SNV annotation                  |       |
|---------------------------------|-------|
| Nonsynonymous                   | 77791 |
| Synonymous                       | 149378|
| Stop gain                       | 640   |
| Stop loss                       | 90    |
| Unknown                          | 5242  |

| DIV annotation                  |       |
|---------------------------------|-------|
| Frameshift deletion             | 638   |
| Frameshift insertion            | 540   |
| Non-frameshift insertion        | 944   |
| Non-frameshift deletion         | 872   |
| Stop gain                       | 29    |
| Stop loss                       | 8     |
| Unknown                          | 129   |
