Genetic variation in the zebrafish

Victor Guryev  
*Netherlands Institute for Developmental Biology*

Marco J. Koudijs  
*Netherlands Institute for Developmental Biology*

Eugene Berezikov  
*Netherlands Institute for Developmental Biology*

Stephen L. Johnson  
*Washington University School of Medicine in St. Louis*

Ronald H.A. Plasterk  
*Netherlands Institute for Developmental Biology*

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

**Recommended Citation**  
https://digitalcommons.wustl.edu/open_access_pubs/2078

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Authors
Victor Guryev, Marco J. Koudijs, Eugene Berezikov, Stephen L. Johnson, Ronald H.A. Plasterk, Fredericus J.M. Van Eeden, and Edwin Cuppen
Genetic variation in the zebrafish

Victor Guryev, Marco J. Koudijs, Eugene Berezikov, et al.

*Genome Res.* 2006 16: 491-497
Access the most recent version at doi:10.1101/gr.4791006

---

**References**

This article cites 26 articles, 15 of which can be accessed free at:
[http://genome.cshlp.org/content/16/4/491.full.html#ref-list-1](http://genome.cshlp.org/content/16/4/491.full.html#ref-list-1)

**Creative Commons License**

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see [http://genome.cshlp.org/site/misc/terms.xhtml](http://genome.cshlp.org/site/misc/terms.xhtml)). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at [http://creativecommons.org/licenses/by-nc/3.0/](http://creativecommons.org/licenses/by-nc/3.0/).

**Email Alerting Service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

---

To subscribe to *Genome Research* go to:
[http://genome.cshlp.org/subscriptions](http://genome.cshlp.org/subscriptions)
Genetic variation in the zebrafish

Victor Guryev,1 Marco J. Koudijs,1 Eugene Berezikov,1 Stephen L. Johnson,2 Ronald H.A. Plasterk,1 Fredericus J.M. van Eeden,1 and Edwin Cuppen1,3

1Hubrecht Laboratory, Netherlands Institute for Developmental Biology, 3584CT, Utrecht, The Netherlands; 2Department of Genetics, Washington University Medical School, St. Louis, Missouri 63130, USA

Although zebrafish was introduced as a laboratory model organism several decades ago and now serves as a primary model for developmental biology, there is only limited data on its genetic variation. An establishment of a dense polymorphism map becomes a requirement for effective linkage analysis and cloning approaches in zebrafish. By comparing ESTs to whole-genome shotgun data, we predicted >50,000 high-quality candidate SNPs covering the zebrafish genome with average resolution of 41 kbp. We experimentally validated ~65% of a randomly sampled subset by genotyping 16 samples from seven commonly used zebrafish strains. The analysis reveals very high nucleotide diversity between zebrafish isolates. Even with the limited number of samples that we genotyped, zebrafish isolates revealed considerable intrastain variation, ranging from 7% (inbred) to 37% (wild-derived) of polymorphic sites being heterozygous. The increased proportion of polymorphic over monomorphic sites results in five times more frequent observation of a three allelic variant compared with human or mouse. Phylogenetic analysis shows that comparisons between even the least divergent strains used in our analysis may provide one informative marker approximately every 500 nucleotides. Furthermore, the number of haplotypes per locus is relatively large, reflecting independent establishment of the different lines from wild isolates. Finally, our results suggest the presence of prominent C-to-U and A-to-I RNA editing events in zebrafish. Overall, the levels and organization of genetic variation between and within commonly used zebrafish strains are markedly different from other laboratory model organisms, which may affect experimental design and interpretation.

The zebrafish (Danio rerio) serves as a unique model for vertebrate development and pharmacological studies (Zon and Peterson 2005). With a draft genome assembly available and thousands of mutants described (Granato and Nusslein-Volhard 1996), a dense map with genetic markers is essential for linkage analysis and cloning approaches. Previous studies employed RAPD (Postlethwait et al. 1994), CA-repeat or simple sequence length polymorphism (SSLP) markers (Shimoda et al. 1999), or single-strand conformational polymorphism (SSCP) markers (Woods et al. 2005) to place >7000 independent markers on various mapping panels. Despite these advances, there is an increasing demand for higher map density that is required for effective positional cloning (Beier 1998).

There are several key advantages that distinguish another type of marker, single nucleotide polymorphism (SNP), as a marker of choice for many genetic studies. To mention a few, SNPs are the most common type of variation in genomes, allowing the generation of ultra-dense genetic maps, and there are efficient low- and high-throughput typing procedures for SNPs currently available (for review, see Vignal et al. 2002). Until now only a low-density SNP-based mapping panel with ~2000 polymorphisms was available (Stickney et al. 2002); the SNP-map contains large gaps up to 58 cM and needs further refinement for routine applications.

In addition to simplifying genetic mapping experiments, studies on genetic variation in model organisms can clarify rate and composition as well as distribution and organization of polymorphic loci in the genome. In particular, it is not clear how much variation still persists in zebrafish laboratory inbred and outbred strains and how it compares to that present in wild isolates. The discovered variation at 9% of tested polymorphic loci in initially homozygous zebrafish C32 strain (Striebeiger et al. 1981) raised a discussion of whether high mutation rate (Buth et al. 1995) or introgression (Nechiporuk et al. 1999) has introduced polymorphisms to this strain. However, the abundance of genetic variation in zebrafish inbred strains alone suggests that individuals within a strain have a diverse genetic background. From this perspective, zebrafish inbred strains differ from other commonly used vertebrate laboratory animals such as inbred mouse or rat strains.

Finally, the analysis of genotype data contributes to better understanding of strain history and the degree of interstrain variation. The variety of methods used to generate inbred lines, e.g., gynogenetic diploids and half-tetrad diploids, inbreeding (for review, see Beier 1998), different natural sources of animals, and breeding regimes, is likely to influence the allele fixation rates in different zebrafish isolates and strains. Knowledge of the phylogenetic relationships between laboratory strains greatly facilitates the choice of strains, which will be most informative for a genetic experiment.

Results and Discussion

Candidate SNP discovery

We have developed a computational SNP discovery pipeline and candidate SNP database named CASCAD (CasCaD SnP Candidate Database).

[The polymorphism and genotype data from this study have been submitted to dbSNP under accession nos. ss49785942–ss49839678.]
and dropouts, meaning that all interpretations based on it should
addition to false duplications also contains other misassemblies
We should mention here that Zv5 is a draft assembly and in
intermediate stage, although a small fraction may result from
redundantly in Zv5 as an artifact of the assembly process in its
nonunique fraction was assigned to fragments that are present
mapped to multiple locations. Presumably, the major part of the
tion on the assembly, and a further 19% of the candidate SNPs
shtml). We failed to place 4% of the candidate SNPs to any loca-
positions in the current zebrafish genome build (Zv5; http://www.
zebrafish genes.

Table 1. Input and output statistics for the computational prediction of zebrafish candidate SNPs using the CASCAD pipeline

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input data (number of reads)</td>
<td></td>
</tr>
<tr>
<td>mRNA</td>
<td>3366</td>
</tr>
<tr>
<td>EST</td>
<td>283,572</td>
</tr>
<tr>
<td>WGS</td>
<td>11,588,394</td>
</tr>
<tr>
<td>Candidate SNPs predicted</td>
<td>51,769</td>
</tr>
<tr>
<td>Synonymous</td>
<td>9111</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>6217</td>
</tr>
<tr>
<td>Nonsense</td>
<td>138</td>
</tr>
</tbody>
</table>

be treated with caution. Our analysis indicates that 73% candid-
date SNPs map to the same linkage group in Zv5 as they would be
placed on gene-based meiotic map of Woods and coworkers
(2005), considerably better than 66% overlap between this mei-
otic map and our candidates mapped on previous Zv4 assembly.

Validation of SNPs
To validate the computationally predicted SNPs, we assayed 398
candidate SNP-containing amplicons evenly distributed over the
25 zebrafish linkage groups (Fig. 1). By resequencing these re-
gions in a panel of 16 individuals representing seven widely used
laboratory strains, we were able to confirm ~65% (256) of them.
This relatively low confirmation rate may at least partially be
explained by the presence of false negatives due to high intra-
strain variation (see below) in combination with the small sam-
ping size per strain (typically two individuals). In addition, the
origins of some of the strains used in EST library construction are
unknown, or samples from the same population were not avail-
able to us. Although segmental genomic duplications could po-
tentially result in false positives, we did not observe evident de-
viations from Hardy-Weinberg equilibrium of allele frequencies,
such as excess of heterozygotes that distinguishes paralogous se-
quence variants from true SNPs, in our verification experiments.

A consequence of validating SNPs by resequencing from ge-
nomic amplicons (average, ~300 bp) was the opportunity to iden-
tify and analyze additional variation. Thus, in addition to the
256 confirmed candidate SNPs, we found as many as 1942 addi-
tional variable positions. Only 155 of these were present in our
database of 51,769 computationally derived SNP candidates. The
high fraction of new SNPs discovered in our validation stage is
accounted for by the presence of intronic and intergenic regions
in our validation assay that could not be scored for polymor-
phisms by our EST and mRNA-centered computational approach.

More than 96% of all polymorphic loci were diallelic (2118/
2198), and the remainder consisted predominantly of short
SSLPs. One-tenth of the variants observed (228) were due to small
insertions or deletions (indels), displaying an intermediate indel
frequency if compared to human and chicken (6.6% and 13.9%,
respectively; source, dbSNP build 124). Only a small fraction of
polymorphisms identified in this study was observed within cod-
ing sequence as annotated in the Ensembl database, with 178 of
them being silent, 85 missense, and two frameshift mutations.

We have designed a Web interface (http://cascad.niob.
knaw.nl/snpview) that facilitates the selection and use of the
validated SNPs in genetic experiments. This tool allows the in-
teractive retrieval and visual representation of validated SNPs for
arbitary combinations of strains.

Candidate SNP characteristics and validation
A comparison of SNP prediction and its verification results for
different organisms can shed light on species-specific character-
istics of variation. Our CASCAD SNP discovery pipeline (Guryev
et al. 2005) used a comparable amount of input data but resulted
in many more candidate polymorphisms for zebrafish than for
rat, providing indirect evidence for higher nucleotide diversity in
zebrafish. A comparison of the verification experiment results in
rat and zebrafish can reveal classes of candidate polymorphisms
with increased or reduced confirmation success rates. We calcu-
lated the correlation between various SNP characteristics and
the confirmation status (Table 2), potentially revealing driving forces.
that shape polymorphism composition in these two organisms. In addition, these correlations were used to define a confirmation likelihood score (categories 0–9), allowing database users to restrict their search to a subset of SNPs with higher expected validation rates.

As expected, candidate SNP verification in both rat and zebrafish is sensitive to the functional context of the polymorphism; silent substitutions are more often verified than are missense. Some trends were found to be species specific: Unlike that in laboratory rat, positive correlation was not found for SNP confirmation at CpG positions in zebrafish. Comparative analysis of methylation and dinucleotide frequencies in different organisms revealed that in spite of the higher methylation level in fish, compared to mammals, CpG depletion is clearly lower in fish (Jabbari and Bernardi 2004). Together with our data, this suggests that CpG possesses more characteristics of a hypervariable site in a mammalian rather than in a fish genome.

Surprisingly, transitive substitutions were less frequently confirmed in zebrafish in contrast to rat, for which they had a higher verification level. As the ratio between transitions and transversions is similar for both organisms, an organism-specific mechanism is suspected. Interestingly, we found two classes of frequently nonconfirmed transitive variants in our verification set, and these correspond to the most frequent type of vertebrate RNA editing events: A<sub>RNA</sub> to G<sub>RNA</sub> (P < 0.1) and C<sub>RNA</sub> to T<sub>DNA</sub> (P < 0.01) due to A-to-I editing and C-to-U editing, respectively. As editing events usually affect multiple consecutively located sites, many of these events may easily be filtered out by our stringent filtering for candidates. Therefore, we performed a computational whole-genome screen for individual mismatches between EST sequences and the zebrafish genome assembly. The incompleteness is exemplified only two samples per strain, but also because of incomplete annotation that these numbers are likely to be an underestimation of the variation data over the complete genome suggests the presence of 425,000 coding SNPs, including 146,000 missense substitutions: The ratio between nonsynonymous and synonymous substitutions per available site (K<sub>a</sub>/K<sub>s</sub>) is 0.142.

Predicted silent substitutions were more frequently confirmed than were missense ones (categories 0–9), allowing database users to restrict their search to a subset of SNPs with higher expected validation rates.

Figure 1. Distribution of candidate and verified SNPs on zebrafish physical map (working draft genome assembly Zv5). Vertical bars represent zebrafish linkage groups; horizontal bars on left side of each linkage group show candidate SNP density given in red for coding and in blue for noncoding candidates (according to Ensembl genome annotation 35.5b, window size = 280 kb). The filled and open boxes to the right of the linkage group correspond to amplicons with confirmed and nonconfirmed candidate SNPs, respectively. The number in each box indicates the total number of confirmed polymorphisms in each amplicon. Genotype information and oligonucleotide primers are available from http://cascad.niob.knaw.nl/snpview.
Nucleotide diversities were calculated as the number of nucleotides scored (bp) divided by the number of SNPs discovered (s). The estimated nucleotide diversity (θ) is given by the equation θ = (s/2k)(∑i=1 to k 1/i), where k is the number of sampled chromosomes. Only 2034 SNPs, out of 2198 in total, are included in this analysis as the others reside in genomic regions that were represented by <50% of the samples.

Table 3. Estimates of nucleotide diversity for different functional fractions of the zebrafish genome

<table>
<thead>
<tr>
<th>Genome fraction</th>
<th>No. of nucleotides scored (bp)</th>
<th>No. of SNPs discovered (s)</th>
<th>Estimated nucleotide diversity (θ)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>20,830</td>
<td>253</td>
<td>3 x 10^-3</td>
</tr>
<tr>
<td>UTR</td>
<td>17,122</td>
<td>333</td>
<td>4.8 x 10^-2</td>
</tr>
<tr>
<td>Introns+noncoding</td>
<td>66,848</td>
<td>1448</td>
<td>5.4 x 10^-2</td>
</tr>
<tr>
<td>Total</td>
<td>104,800</td>
<td>2034</td>
<td>4.8 x 10^-3</td>
</tr>
</tbody>
</table>

* Nucleotide diversities were calculated as θ = (s/2k)(∑i=1 to k 1/i), where k is the number of sampled chromosomes. Only 2034 SNPs, out of 2198 in total, are included in this analysis as the others reside in genomic regions that were represented by <50% of the samples.

Figure 2. Neighbor-joining tree for 16 zebrafish samples representing seven different strains, based on genotyping of 2120 SNPs. Coefficients represent bootstrap test support values for tree nodes.
be ample for the selection of sufficient SNP markers. For example, the rate of polymorphisms homozygous in both closely related AB and Tu strains is estimated to be about one per 500 bp.

Intrastrain variation
Most zebrafish lines originate and are maintained as outbred stocks. Only C32 and SJD have been bred to obtain inbred lines. Although most strains are kept as independent stocks at many laboratories worldwide, only very limited data are available on the degree of genetic variation within a line and the potential genetic differences between various (sub-)stocks. As expected, the Singapore local wild-type isolate \((n = 4)\) was found to be the most heterozygous “strain,” with 37% of the SNPs being polymorphic; 14.1%, 14.6%, 17.6%, and 24.8% of the SNPs are heterozygous in WIK, Tu, TL, and AB, respectively \((n = 2\) per strain). For the inbred strains, we found that 7% and 11% of the loci are polymorphic in SJD and C32, respectively, which is in line with previous observations showing that inbred zebrafish strains are not genetically uniform (Buth et al. 1995; Nechiporuk et al. 1999). Interestingly, most of the heterozygous loci \((172/184)\) in the C32 strain are also polymorphic in the other samples, supporting the hypothesis that these polymorphisms originate from a common origin, were inherited, and did not appear in this strain due to mutation process as was proposed earlier (Buth et al. 1995).

Structure of genetic variation
An important question for any model organism is organization of its genetic variation. A limited number of founder animals and continuous inbreeding result in genome blocks with limited haplotype diversity that can greatly simplify genetic and QTL mapping in laboratory strains. Data available for eight mouse laboratory strains (Yalcin et al. 2004), show that most multi-SNP regions resequenced \((136/226)\); average, 5.32 SNPs/region) contain only two haplotypes. On the opposite, only a small fraction of zebrafish amplicons containing multiple SNPs \((19/334)\); average, 5.74 SNPs/amplicon; Singapore wild-types were excluded from the analysis) is compatible with the presence of only two haplotypes, suggesting that zebrafish strains used in a laboratory will not reveal a pronounced high-level structure of genomic variation, providing little reason for building a detailed haplotype map for this organism.

Conclusions
The degree and organization of genetic variation between and within zebrafish strains was not found to be comparable to other commonly used vertebrate model organisms. Therefore, one should take into account possible effects of genetic variation in experimental design and interpretation and should be careful when comparing results from different laboratories using different (sub-)strains. The development and use of well-characterized inbred zebrafish lines, preferentially marked with a unique recessive phenotype, could significantly reduce confounding effects resulting from genetic heterogeneity.

Methods
SNP discovery
The mRNA and EST sequence data used in this study were downloaded from NCBI GenBank \((http://www.ncbi.nlm.nih.gov/Genbank)\) and Ensembl trace repository \((http://trace.ensembl.org)\). EST sequences and quality data from Singapore isolate were provided by Dr. Jinrong Peng \((Institute of Molecular and Cell Biology, Singapore)\). We used Ensembl trace archive \((http://trace.ensembl.org)\) as a source of genomic traces. EST and mRNA sequences were masked for zebrafish-specific repeats, low-complexity regions, and zebrafish mitochondrial DNA by using RepeatMasker. Local SSAHA searches were performed to collect hits with nearly exact homology containing a single mismatch in mRNA/EST subset and remote searches (using Ensembl SSAHA search server) in case of mRNA/EST versus WGS comparison. Only hits with a high-quality mismatch \((phred score >20 for both reads)\) within a sequence stretch of >80-bp identity were retained. The mRNA subset that is not annotated for base-calling quality data was treated as having a reliable overall quality. Hits were clustered to represent unique variations and stored in a MySQL database. Candidate SNPs were annotated and placed on the Zv5 genome assembly by using methods reported previously (Guryev et al. 2004).

Predicted and discovered SNPs as well as genotype data obtained in this study were submitted to dbSNP under the following accession numbers: ss49785942–ss49839678. The CASCAD database of candidate SNPs and underlying supporting information is publicly available at \(http://cascad.niob.knaw.nl\). All scripts are freely available upon request.

SNP validation
For the verification experiment, we used 16 samples from seven different zebrafish isolates: AB \((two individuals)\), C32 \((two)\), SJD \((two)\), TL \((two)\), Tu \((two)\), WIK \((two)\), and Singapore wild type \((four)\). AB, TL, Tu, and WIK samples were taken from the colony kept at the Hubrecht Laboratory, C32 and SJD originated from Washington University, and the Singapore wild types were kindly provided by Dr. Jinrong Peng. DNA isolation was done by using the protocol described in Westerfield (2000).

We have semirandomly sampled candidate SNPs to generate a set of markers with even distribution throughout the zebrafish linkage groups. For this purpose we have divided the assembled zebrafish genome into equally sized bins and randomly selected a candidate from each bin. Primers for PCR amplification and sequencing of the genomic region were designed by using a customized Web interface \((http://primers.niob.knaw.nl)\) to the Primer3 program \((http://www-genome.wi.mit.edu/genome_software/other/primer3.html)\). Primer sequences can be obtained upon request or retrieved interactively from the Web interface \((http://cascad.niob.knaw.nl/snpview)\) that allows the retrieval and visual representation of validated SNPs between arbitrary combinations of strains.

PCRs were carried out by using a touchdown thermocycling program \((60 sec at 92°C; 30 cycles for 20 sec at 92°C, 20 sec at 65°C with a decrement of 0.4°C per cycle, and 30 sec at 72°C; followed by 10 cycles of 20 sec at 92°C, 20 sec at 58°C, and 30 sec at 72°C; and 18 sec at 72°C; GeneAmp9700, Applied Biosystems)\) and contained 30–50 ng genomic DNA, 0.2 µM of each forward primer and 0.2 µM of each reverse primer, 400 µM each of dNTP, 25 mM Tricine, 7.0% glycerol \((w/v)\), 1.6% DMSO \((w/v)\), 2 mM MgCl\(_2\), 85 mM ammonium acetate \((pH 8.7)\), and 0.2 U Taq polymerase in a total volume of 10 µL. After thermocycling, the PCR reactions were diluted with 25 µL water and mixed by pipetting, and 1 µL was used as template for dideoxy cycle sequencing, as recommended by the manufacturer \((BigDye v3.1, Applied Biosystems)\) using one of the primers used for the PCR amplification. Sequencing reactions were analyzed on an ABI3730XL capillary sequencer \((Applied Biosystems)\), and the obtained sequences were scored for polymorphic positions by using the PolyPhred program \((Nickerson et al. 1997)\) followed by manual inspection.
Phylogenetic reconstruction

Sequence alignments of 2120 confirmed variable positions were used as an input for the MEGA3 program (Kumar et al. 2004). Phylogenetic tree was built with a Neighbor-joining algorithm using p-distances with a pair-wise deletion option. Support for each node was determined by a bootstrap test.

Whole-genome mutation-type screen

EST sequences were mapped to zebrafish assembly Zv4 by using the GMAP program (Wu and Watanabe 2005). We scored only candidate SNPs occurring in exons having at least 90% identity between EST and genome sequences. Genome annotation from Ensembl build 31.4d was used to deduce alleles observed in genomic and cDNA-based reads.

Acknowledgments

We thank Dr. Jinrong Peng (Institute of Molecular and Cell Biology, Singapore), Washington University St. Louis, and AgenCourt Bioscience Corporation for providing zebrafish EST sequence and/or quality data, and the Zebrafish Sequencing Group at the Wellcome Trust Sanger Institute for making the WGS trace data and zebrafish genome assembles publically available before publication. This work was supported by NWO genomics grant 050-10-024.

References


Received October 10, 2005; accepted in revised form January 18, 2006.