2006

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

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

Letter

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 Cuppen[1,3]

[1]*Hubrecht Laboratory, Netherlands Institute for Developmental Biology, 3584CT, Utrecht, The Netherlands;* [2]*Department 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 interstrain 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 polymorphism and genotype data from this study have been submitted to dbSNP under accession nos. ss49785942–ss49839678.]

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 (Streisinger 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

[3]Corresponding author.
E-mail ecuppen@niob.knaw.nl; fax 31-30-251-6554.
Article published online ahead of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/ gr.4791006.

Database, http://cascad.niob.knaw.nl) (Guryev et al. 2004, 2005). Here, we report an application of this pipeline for construction and verification of a single-nucleotide variation database for zebrafish (*D. rerio*). Publicly available EST and mRNA sequences were compared to each other and to whole-genome shotgun (WGS) trace sequences for the occurrence of SNPs (Table 1). We did not predict variations within the WGS set. The polymorphisms that are observed in expressed sequences may be particularly useful for genetic mapping studies as well as for drawing inferences about functional differences of genes in different strains.

The resulting raw data set (>1 × 10^6 mismatches) was filtered for high-ranking candidate SNPs based on a variety of parameters, including masking out repetitive sequences and the presence of high PHRED quality score (>20) in each of the candidate SNP alleles. After clustering, 51,769 unique candidate SNPs were obtained. Although a similar amount of input EST sequences was used in this study compared with our previous analysis for the rat (Guryev et al. 2004), we found 55% more candidate SNPs, suggesting a higher SNP frequency in zebrafish. Despite the high number of SNPs discovered in this effort, the CASCAD database currently contains only a small part of variation associated with expressed sequences. This point is illustrated by the small overlap of the CASCAD SNP set with the previously published set of EST-derived SNPs (Stickney et al. 2002), of which only 6% are present in our database. Furthermore, intergenic regions and introns are expected to harbor polymorphisms with much higher density compared with expressed sequences that possess functional constraints such as coding capacity and splice signals.

The average frequency of candidate SNP is 1 per 41 kbp, and the largest gap between two adjacent markers is 2 Mb on linkage group 14. About one-third of the candidates reside in genomic regions that are annotated as protein coding, including 9111 synonymous and 6375 nonsynonymous changes. The candidate SNPs cover 13,016 of 31,219 UniGene clusters and 7841 of 22,877 predicted Ensembl genes, or approximately one-third of zebrafish genes.

Over 66% of the candidate SNPs could be assigned to unique positions in the current zebrafish genome build (Zv5; http://www.sanger.ac.uk/Projects/D_rerio/ Zv5_assembly_information.shtml). We failed to place 4% of the candidate SNPs to any location on the assembly, and a further 19% of the candidate SNPs mapped to multiple locations. Presumably, the major part of the nonunique fraction was assigned to fragments that are present redundantly in Zv5 as an artifact of the assembly process in its intermediate stage, although a small fraction may result from sequence difference between otherwise highly similar paralogs. We should mention here that Zv5 is a draft assembly and in addition to false duplications also contains other misassemblies and dropouts, meaning that all interpretations based on it should be treated with caution. Our analysis indicates that 73% candidate 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 meiotic map and our candidates mapped on previous Zv4 assembly.

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

Input data (number of reads)	
mRNA	3366
EST	283,572
WGS	11,588,394
Candidate SNPs predicted	51,769
Synonymous	9111
Nonsynonymous	6217
Nonsense	158

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 regions 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 sampling 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 available to us. Although segmental genomic duplications could potentially result in false positives, we did not observe evident deviations from Hardy-Weinberg equilibrium of allele frequencies, such as excess of heterozygotes that distinguishes paralogous sequence variants from true SNPs, in our verification experiments.

A consequence of validating SNPs by resequencing from genomic amplicons (average, ~300 bp) was the opportunity to identify and analyze additional variation. Thus, in addition to the 256 confirmed candidate SNPs, we found as many as 1942 additional 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 polymorphisms 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 coding 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 interactive retrieval and visual representation of validated SNPs for arbitrary 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 characteristics 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 calculated the correlation between various SNP characteristics and the confirmation status (Table 2), potentially revealing driving forces

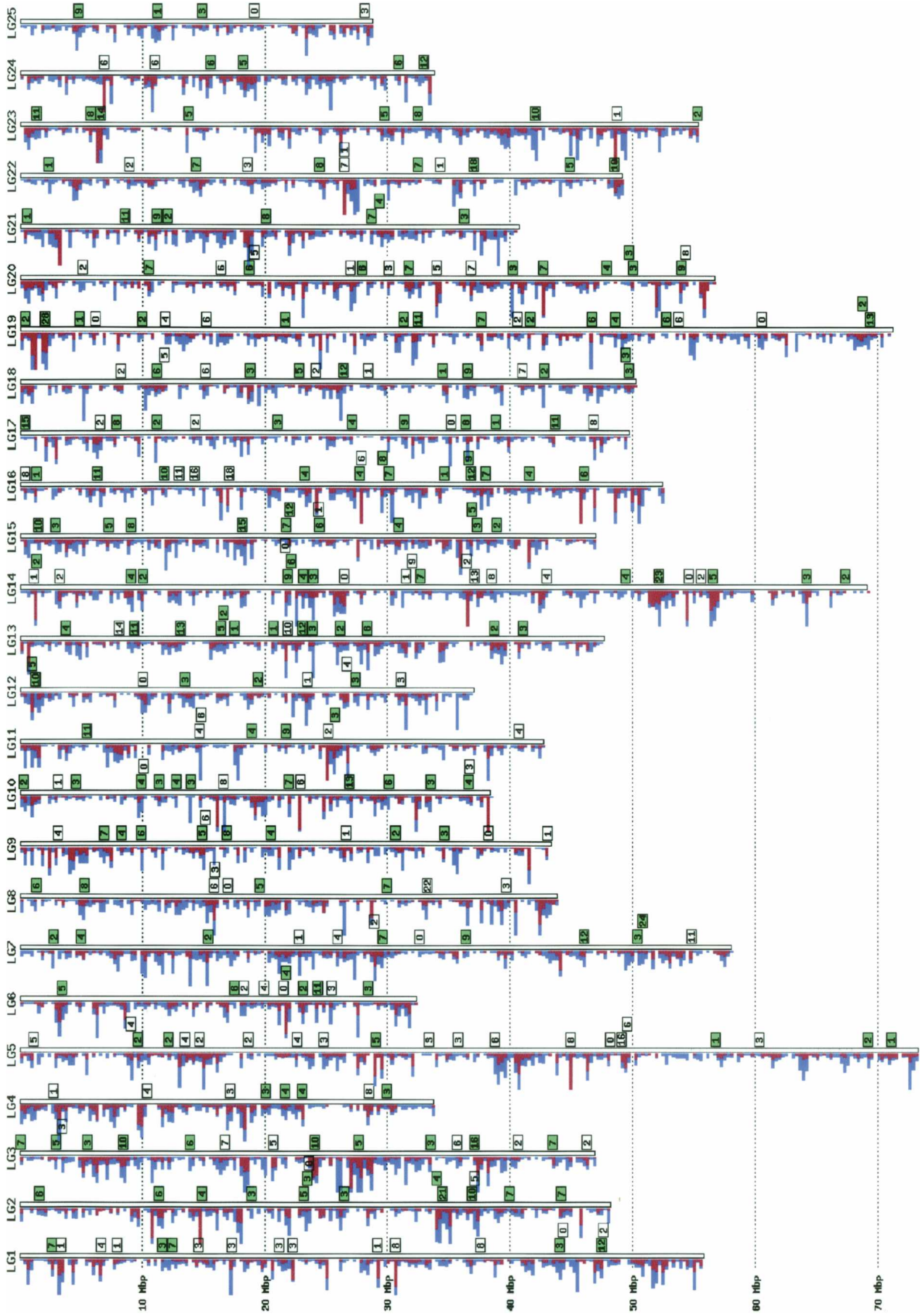

Figure 1. (Legend on next page).

Table 2. Correlation between validation of candidate SNPs and their characteristics

Characteristics	Correlation coefficient[a]	Correlation description
SNP functional class	−0.1680	Predicted silent substitutions were more frequently confirmed than were missense ones
EST vs. EST	−0.2358	Candidate SNPs observed between EST reads were less reliable
EST vs. WGS	0.1618	Variants observed between EST and WGS reads were more likely to be confirmed
Number of reads	0.1221	Candidates supported by multiple reads for each allele had increased verification success
CpG	0.0017[b]	Polymorphisms predicted at hypervariable CpG sites were more frequently confirmed
Transition	−0.1311	Transitive candidates (G↔A, T↔C) were less reliable than were transversive ones

[a]Pearson r correlation coefficient based on 398 samples.
[b]$P > 0.05$, not significant.

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_{DNA} to G_{cDNA} ($P < 0.1$) and C_{DNA} to T_{cDNA} ($P < 0.01$) 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 search was restricted to the sense strand as annotated in Ensembl build 31 and showed 8% overrepresentation of A-to-G over G-to-A substitutions and 11% excess of C-to-T versus T-to-C substitutions. From the analysis of this limited set of ESTs, we estimate that there are at least 2600 editing sites (C-to-U and A-to-I). Similarly to primates, RNA editing may be very abundant in zebrafish, with a frequency of A-to-I editing of one order of magnitude larger compared with that of mouse, rat, chicken, or fly (Eisenberg et al. 2005). However in contrast to primates, C-to-U editing events seem to be more common than are A-to-I editing in zebrafish.

When we now eliminate all nonconfirmed polymorphisms from our confirmation experiment that may have been due to RNA-editing events, we observe a positive, although not significant (possibly due to lower sample size, $n = 339$) verification correlation with both CpG sites and transitive mutations, similar as for the rat. These results strongly suggest that high rates RNA editing events in zebrafish account for the observed relatively low confirmation rate of transitive candidate SNPs. We need to note that in absence of solid experimental data, one cannot exclude an alternative explanation for this apparent bias between ESTs and genomic sequence, namely, the occurrence of cytosine deamination during sample preparation and library construction, but it seems unlikely as it is observed for two independent EST data sets (Washington University EST project, http://genome.wustl.edu/est) (Lo et al. 2003).

Nucleotide diversity

Our validation assay employed ~400 amplicons evenly distributed throughout all zebrafish linkage groups and covering coding, intronic, untranslated, and intergenic parts of its genome with the great prevalence of newly discovered variants over computationally predicted SNPs. This justifies the use of our genotyping results for establishing an estimate of nucleotide diversity in zebrafish. To this end, we used only regions for which high-quality sequence data were obtained for at least half of the samples that were tested (total, 105 kbp). Gene annotation from the Ensembl zebrafish genome database build 33 enabled us to calculate the nucleotide diversity for functionally different fractions of the genome (Table 3). Polymorphisms are distributed in a nonrandom fashion between coding and noncoding parts of the genome ($P < 0.001$) at a rate of one per 82 and 47 bp, respectively. Within coding sequences, there is a strong bias toward synonymous substitutions: The ratio between nonsynonymous and synonymous substitutions per available site (K_a/K_s) is 0.142. This value is similar in other vertebrate organisms and illustrates a pronounced negative selection on coding regions. Extrapolation of the variation data over the complete genome suggests the presence of 425,000 coding SNPs, including 146,000 missense variants that may result in phenotypic effects. It should be mentioned that these numbers are likely to be an underestimation because some polymorphisms that were missed in our study assayed only two samples per strain, but also because of incomplete genome assembly and annotation. The incompleteness is exem-

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 kbp). 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.

plified by the fact that although all candidate regions are associated with an EST read, only 206, roughly half of the tested amplicons, correspond to an annotated Ensembl transcript (based on zebrafish mRNAs and ESTs), while 57 of them overlap with predicted transcripts (based on ESTs and predicted open reading frames) and the remaining 135 were not associated with any known or predicted transcript.

Similarly, the estimated average nucleotide diversity (Table 3) is likely to be an underestimate as there is a bias toward functionally constrained expressed sequences in our verification set. Strikingly, even this value is about one order of magnitude higher than that observed in human populations (Deutsch et al. 2001) and four times higher than that in a large set of commonly used rat strains (Guryev et al. 2004) or that between two mouse subspecies (Wade et al. 2002). Although this value is similar to that observed in three *Drosophila* species (Moriyama and Powell 1996), it represents, to our knowledge, the highest nucleotide diversity seen in any vertebrate species. Since our estimate is based on genome-wide selection of SNP markers and thus reflects abundance of polymorphic loci in the zebrafish genome, and given the fact that teleosts make up more than half of living vertebrate species (23,600 species) (Helfman et al. 1997), high nucleotide diversity may be a prerequisite factor enabling the rapid radiation that is commonly seen in teleosts. An increased variation at the nucleotide level could comprise a genetic basis for phenotypic differences underlying adaptive evolution and, together with recent whole-genome duplication events, promoting speciation. However, SNP data from more teleosts will be needed to conclude if the high nucleotide diversity is specific for zebrafish or common to teleosts.

The high nucleotide diversity in zebrafish also results in more frequent occurrence of three alleles at a single locus. About 1% of single nucleotide variants had three alleles, which is significantly higher than observed in mouse, human, or chicken (0.19%, 0.22%, and 0.28%, respectively, as calculated from NCBI dbSNP build 124). The observed number of triallelic SNPs in zebrafish is close to an estimate based on diallelic SNPs frequency (18–21), suggesting that most triallelic SNPs result from independent, unselected mutations, rather than the identification of sites of strong positive selection. Although such triallelic SNPs are mostly neglected in genetic studies in other vertebrates where they are rare, in zebrafish they may prove advantageous in designing mapping probes sets useful for a greater fraction of loci tested across a wider variety of genetic backgrounds used.

Phylogenetic relationships

To assess the degree of similarity between different strains, we used 2120 confirmed SNPs that were genotyped in 16 zebrafish samples representing seven widely used laboratory isolates for construction of a phylogenetic tree (Fig. 2). The tree is mostly consistent with previously described CA-repeat and SSCP-based

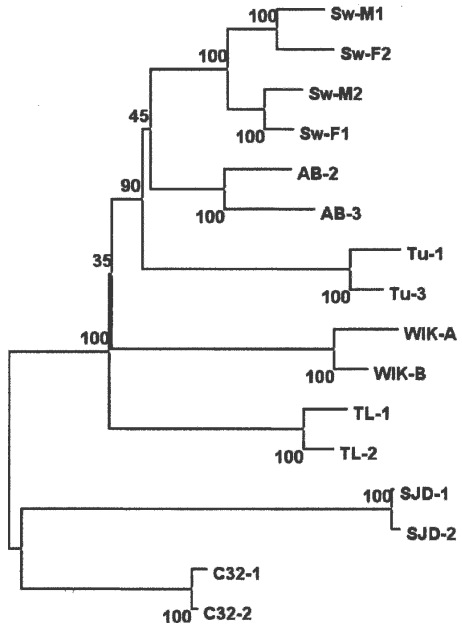

Figure 2. Neighbor-joining tree for 16 zebrafish samples representing seven different strains, based on genotyping of 2120 SNPs. Coefficients represent bootstrap test support values for tree nodes.

results (Knapik et al. 1998; Nechiporuk et al. 1999). The relatively large distance between most of the lines can be explained by the high nucleotide diversity in combination with the recent independent establishment of different laboratory lines from wild populations from around the world (Trevarrow and Robison 2004). This is fundamentally different from the history of laboratory mice and rats, which are thought to have originated from a limited source of domesticated animals (Beck et al. 2000; Hedrich 2000).

The SJD and C32 strains are the most polymorphic with respect to any of the other strains that we analyzed due in part to fixation to homozygosity of many unique alleles in these inbred strains. Nevertheless, the closest relationship of our C32 isolate with SJD contradicts with the previously observed lowest divergence between C32 and AB (Nechiporuk et al. 1999) and may reflect the more recent breeding and maintenance history of these lines, by which genes from SJD were introgressed into C32, to enhance strain vigor, and reciprocally, genes from C32 were introgressed into SJD to improve sex ratios. Resultant C32 and SJD lines each bear ~5%–10% of SSLP markers from the reciprocal line (Rawls et al. 2003).

Although this phylogenetic tree can be used to choose optimal pairs of strains for setting up genetic mapping experiments, the diversity between any of the lines will in most cases already

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

Genome fraction	No. of nucleotides scored (bp)	No. of SNPs discovered (s)	Estimated nucleotide diversity (θ)[a]
Coding	20,830	253	3×10^{-3}
UTR	17,122	333	4.8×10^{-3}
Introns+noncoding	66,848	1448	5.4×10^{-3}
Total	104,800	2034	4.8×10^{-3}

[a]Nucleotide diversities were calculated as $\theta = (s/n\sum_{i=1}^{k-1}(1/i))$, 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.

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 every 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 of each 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 assemblies publically available before publication. This work was supported by NWO genomics grant 050-10-024.

References

Beck, J.A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J.T., Festing, M.F., and Fisher, E.M. 2000. Genealogies of mouse inbred strains. *Nat. Genet.* **24:** 23–25.

Beier, D.R. 1998. Zebrafish: Genomics on the fast track. *Genome Res.* **8:** 9–17.

Buth, D.G., Gordon, M.S., Plaut, I., Drill, S.L., and Adams, L.G. 1995. Genetic heterogeneity in isogenic homozygous clonal zebrafish. *Proc. Natl. Acad. Sci.* **92:** 12367–12369.

Deutsch, S., Iseli, C., Bucher, P., Antonarakis, S.E., and Scott, H.S. 2001. A cSNP map and database for human chromosome 21. *Genome Res.* **11:** 300–307.

Eisenberg, E., Nemzer, S., Kinar, Y., Sorek, R., Rechavi, G., and Levanon, E. 2005. Is abundant A-to-I editing primate-specific? *Trends Genet.* **21:** 77–81.

Granato, M. and Nusslein-Volhard, C. 1996. Fishing for genes controlling development. *Curr. Opin. Genet. Dev.* **6:** 461–468.

Guryev, V., Berezikov, E., Malik, E., Plasterk, R.H.A., and Cuppen, E. 2004. Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Res.* **14:** 1438–1443.

Guryev, V., Berezikov, E., and Cuppen, E. 2005. CASCAD: A database of annotated single nucleotide polymorphisms associated with expressed sequences. *BMC Genomics* **6:** 10.

Hedrich, H.J. 2000. History, strains and models. In: *The laboratory rat: The handbook of experimental animals* (ed. G.J. Krinke), pp. 3–16. Academic Press, NY.

Helfman, G.S., Colette, B.B., and Facey, D.E. 1997. *The diversity of fishes.* Blackwell Science, Malden, MA.

Jabbari, K. and Bernardi, G. 2004. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* **333:** 143–149.

Knapik, E.W., Goodman, A., Ekker, M., Chevrette, M., Delgado, J., Neuhauss, S., Shimoda, N., Driever, W., Fishman, M.C., and Jacob, H.J. 1998. A microsatellite genetic linkage map for zebrafish (*Danio rerio*). *Nat. Genet.* **18:** 338–343.

Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform.* **5:** 150–163.

Lo, J., Lee, S., Xu, M., Liu, F., Ruan, H., Eun, A., He, Y., Ma, W., Wang, W., Wen, Z., et al. 2003. 15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Res.* **13:** 455–466.

Moriyama, E.N. and Powell, F.R. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13:** 261–277.

Nechiporuk, A., Finney, J.E., Keating, M.T., and Johnson, S.L. 1999. Assessment of polymorphism in zebrafish mapping strains. *Genome Res.* **9:** 1231–1238.

Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. Polyphred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25:** 2745–2751.

Postlethwait, J., Johnson, S., Midson, C., Talbot, W., Gates, M., Ballinger, E., Africa, D., Andrews, R., Carl, T., and Eisen, J. 1994. A genetic linkage map for the zebrafish. *Science* **264:** 699–703.

Rawls, J.F., Frieda, M.R., McAdow, A.R., Gross, J.P., Clayton, C.M., Heyen, C.K., and Johnson, S.L. 2003. Coupled mutagenesis screens and genetic mapping in zebrafish. *Genetics* **163:** 997–1009.

Shimoda, N., Knapik, E.W., Ziniti, J., Sim, C., Yamada, E., Kaplan, S., Jackson, D., de Sauvage, F., Jacob, H., and Fishman, M.C. 1999. Zebrafish genetic map with 2000 microsatellite markers. *Genomics* **58:** 219–232.

Stickney, H.L., Schmutz, J., Woods, I.G., Holtzer, C.C., Dickson, M.C., Kelly, P.D., Myers, R.M., and Talbot, W.S. 2002. Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays. *Genome Res.* **12:** 1929–1934.

Streisinger, G., Walker, C., Dower, N., Knauber, D., and Singer, F. 1981. Production of clones of homozygous diploid zebra fish (*Brachydanio rerio*). *Nature* **291:** 293–296.

Trevarrow, B. and Robison, B. 2004. Genetic backgrounds, standard lines, and their husbandry. In: *The zebrafish: Cellular and developmental biology, genetics, genomics and informatics* (eds. H.W. Detrich III, et al.), pp. 599–616. Academic Press, NY.

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* **34:** 275–305.

Wade, C.M., Kulbokas III, E.J., Kirby, A.W., Zody, M.C., Mullikin, J.C., Lander, E.S., Lindblad-Toh, K., and Daly, M.J. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature* **420:** 574–578.

Westerfield, M. 2000. *The zebrafish book: A guide for the laboratory use of zebrafish* (Danio rerio), 4th ed. University of Oregon Press, Eugene, OR.

Woods, I.G., Wilson, C., Friedlander, B., Chang, P., Reyes, D.K., Nix, R., Kelly, P.D., Chu, F., Postlethwait, J.H., and Talbot, W.S. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res.* **15:** 1307–1314.

Wu, T.D. and Watanabe, C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21:** 1859–1875.

Yalcin, B., Fullerton, S., Miller, S., Keays, D.A., Brady, S., Bhorma, A., Jefferson, A., Volpi, E., Copley, R.R., and Flint, J. 2004. Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc. Natl. Acad. Sci.* **101:** 9734–9739.

Zon, L.I. and Peterson, R.T. 2005. In vivo drug discovery in the zebrafish. *Nat. Rev. Drug Discov.* **4:** 35–44.

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