Title

Annotation of the Zebrafish Genome through an Integrated Transcriptomic and Proteomic Analysis

Journal

Molecular & Cellular Proteomics, 13(11)

ISSN

1535-9476

Authors

Kelkar, Dhanashree S
Provost, Elayne
Chaerkady, Raghothama
et al.

Publication Date

2014-11-01

DOI

10.1074/mcp.m114.038299

Peer reviewed

Annotation of the Zebrafish Genome through an Integrated Transcriptomic and Proteomic Analysis

Dhanashree S. Kelkar*‡, Elayne Provost§, Raghothama Chaerkady¶, Babylakshmi Muthusamy*‖, Srikanth S. Manda*‖, Tejaswini Subbannayya*‡, Lakshmi Dhevi N. Selvan*‡, Chieh-Huei Wang¶, Keshava K. Datta*‡‡, Sunghee Woo§§, Sutopa B. Dwivedi*‡, Santosh Renuse*‡, Derese Getnet¶, Tai-Chung Huang¶, Min-Sik Kim¶**, Sneha M. Pinto*¶¶¶, Christopher J. Mitchell¶, Anil K. Madugundu*, Praveen Kumar*, Jyoti Sharma*¶¶, Jayshree Advani*, Gourav Dey*¶¶, Lavanya Balakrishnan*‖‖, Nazia Syed*[a], Vishalakshi Nanjappa*‡, Yashwanth Subbannayya*, Renu Goel*, T. S. Keshava Prasad*‡‖¶¶, Vineet Bafna§§, Ravi Sirdeshmukh*, Harsha Gowda*, Charles Wang[bc], Steven D. Leach§¶[c], and Akhilesh Pandey*¶**[cde]**

Accurate annotation of protein-coding genes is one of the primary tasks upon the completion of whole genome sequencing of any organism. In this study, we used an integrated transcriptomic and proteomic strategy to validate and improve the existing zebrafish genome annotation. We undertook high-resolution mass-spectrometry-based proteomic profiling of 10 adult organs, whole adult fish body, and two developmental stages of zebrafish (SAT line), in addition to transcriptomic profiling of six organs. More than 7,000 proteins were identified from proteomic analyses, and ~69,000 high-confidence transcripts were assembled from the RNA sequencing data. Approximately 15% of the transcripts mapped to intergenic regions, the majority of which are likely long non-coding RNAs. These high-quality transcriptomic and proteomic data were used to manually reannotate the zebrafish genome. We report the identification of 157 novel protein-coding genes. In addition, our data led to modification of existing gene structures including novel exons, changes in exon coordinates, changes in frame of translation, translation in annotated UTRs, and joining of genes. Finally, we discovered four instances of genome assembly errors that were supported by both proteomic and transcriptomic data. Our study shows how an integrative analysis of the transcriptome and the proteome can extend our understanding of even well-annotated genomes. ***Molecular & Cellular Proteomics 13: 10.1074/mcp.M114.038299, 3184–3198, 2014.***

From the *Institute of Bioinformatics, International Technology Park, Bangalore 560 066, India; ‡Amrita School of Biotechnology, Amrita University, Kollam 690 525, India; §Department of Surgery, Johns Hopkins University, Baltimore, Maryland 21205; ¶McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland 21205; ‖Centre of Excellence in Bioinformatics, School of Life Sciences, Pondicherry University, Puducherry 605014, India; **Departments of Biological Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205; ‡‡School of Biotechnology, KIIT University, Bhubaneswar, Odisha 751024, India; §§Department of Computer Science, University of California, San Diego, California 92093; ¶¶Manipal University, Madhav Nagar, Manipal, Karnataka 576104, India; ‖‖Department of Biotechnology, Kuvempu University, Shimoga 577 451, India; [a]Department of Biochemistry and Molecular Biology, School of Life Sciences, Pondicherry University, Puducherry 605 014, India; [b]The Center for Genomics and Division of Microbiology & Molecular Genetics, School of Medicine, Loma Linda University, Loma Linda, California 92350; [d]Sol Goldman Pancreatic Cancer Research Center, Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205; [e]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

Received February 6, 2014, and in revised form, July 11, 2014
Published, MCP Papers in Press, July 24, 2014, DOI 10.1074/mcp.M114.038299

Author contributions: D.S.K., C.W., S.D.L., and A.P. designed research; D.S.K., E.P., R.C., L.N.S., C.-H.W., K.K.D., S.B.D., S.R., D.G., T.H., M.K., S.M.P., C.W., and A.P. performed research; C.W., S.D.L., and A.P. contributed new reagents or analytic tools; D.S.K., B.M., S.S.M., T.S., S.W., S.B.D., S.M.P., C.J.M., A.K.M., P.K., J.S., J.A., L.B., N.S., V.N., Y.S., R.G., and V.B. analyzed data; D.S.K. and A.P. wrote the paper; G.D. provided figure illustrations; T.P., V.B., R.S., H.G., C.W., S.D.L., and A.P. provided critical comments.

Zebrafish (*Danio rerio*) is an important vertebrate model organism that has been widely used in biomedical research in several areas, including developmental biology, disease biology, toxicology, and behavior. The latest genome assembly, Zv9, which was released in October 2011, combines the advantages of clone-by-clone sequencing and shotgun sequencing technologies. In this assembly, 83% of sequences were generated from capillary sequencing of clones, with the

gaps filled by shotgun sequencing reads generated via next-generation sequencing (1). The genome assembly includes 25 chromosomes along with 995 contigs from shotgun sequencing that could not be assembled into chromosomes. The extent and quality of the genome annotation ultimately determine the usefulness of the genome sequence itself. The current Ensembl genome annotation set (genebuild release 75) has 56,754 transcripts corresponding to 33,737 genes. This gene set includes annotations from an automated Ensembl annotation pipeline, a VEGA manual annotation pipeline (VEGA Release 55), and transcript models derived from RNA-Seq-derived[1] data from five adult tissues and seven developmental stages of zebrafish (2).

Shotgun proteomics and NextGen sequencing data have great potential to assist in genome annotation through automated as well as manual strategies. With advancements in the methods for transcriptome profiling and data processing, an increasing number of studies are being carried out in which transcriptomic and proteomic data are analyzed in an integrative manner (3, 4). There are also a number of reports of identification of novel coding loci using transcriptomic data alone or in combination with proteomic data (5, 6). Our previous efforts have successfully demonstrated the power of proteogenomic analyses in improving genome annotation, as exemplified by studies on *Mycobacterium tuberculosis*, *Candida glabrata*, *Leishmania donovani*, *Anopheles gambiae*, and *Homo sapiens* (7–11).

Here, we report the use of in-depth transcriptomic and proteomic profiling to refine the genome annotation of zebrafish (Fig. 1). The transcriptomic (RNA-Seq) data were derived from six adult organs. We identified 69,206 high-confidence transcripts, including novel transcripts for 22,585 genes and 9,404 novel transcribed loci. In total, 6,975 proteins were identified via proteomic analysis of 10 different adult organs, whole adult fish body, and two developmental stages. We employed various proteogenomic strategies that included searching the mass spectra against a number of custom databases, including a six-frame translated genome database, a translated RNA-Seq transcript database, and a *de novo* gene prediction set. To reduce false positives (12, 13), we manually verified the peptide spectrum matches (PSMs) identified from each of these searches. Novel peptides obtained from only good-quality spectral matches were considered for genome annotation improvement. Apart from the identification of novel genes, significant findings of our study include the identification of genome assembly errors, novel exons, novel splice forms, and alternate translational start sites.

[1] The abbreviations used are: RNA-Seq, RNA sequencing; PSM, peptide spectrum match; CPAT, Coding-Potential Assessment Tool; FDR, false discovery rate.

EXPERIMENTAL PROCEDURES

Construction of Libraries and RNA-Seq Analysis—A genetically defined Zebrafish SAT line (Sanger AB Tübingen) was procured and cultured in an in-house fish facility. Muscle, liver, intestine/pancreas, testis, eye, and spleen were dissected and collected in RNA*later* on ice before RNA extraction. Total RNA was isolated from each organ using a Qiagen RNeasy Kit (Qiagen, Inc., Carlsbad, CA) according to the manufacturer's protocol. RNA-Seq of these six organs/tissues was performed according to the manufacturer's protocol using the Illumina TruSeq RNA Sample Preparation Kit and SBS Kit v3 (Illumina, San Diego, CA). Briefly, RNA quality was determined using an Agilent Bioanalyzer with an RNA Nano 6000 chip. RNA-Seq library construction was started using 500 ng of total RNA that was then subjected to poly(A)$^+$ selection and fragmentation. Followed by first and second strand synthesis, the cDNA was subjected to end repair, adenylation of 3′ ends, and adapter ligation. One of six unique indices was used in each individual sample. After AMPure XP magnetic bead (Beckman Coulter, Brea, CA) clean-up, each cDNA sample was subjected to 15 cycles of PCR amplification using an ABI 9700 thermal cycler. The cDNA library quality and size distribution were checked using an Agilent Bioanalyzer with a DNA 1000 chip. Our libraries showed a size between 200 and 500 bp with a peak at ~260 bp. All libraries were carefully quantitated using a Qubit 2.0 fluorometer (Invitrogen, Grand Island, NY) and were stored in microfuge tubes (Invitrogen) in a −20 °C freezer. The cluster generation was done using an Illumina TruSeq V3 flow cell with six different cDNA libraries with different indices in each lane, repeated in three lanes, at a concentration of ~8.6 pM. RNA-Seq was carried out on Illumina's HiScanSQ system (Illumina) using the Illumina TruSeq SBS V3 sequencing kit and 50 bp by 50 bp paired reads.

RNA-Seq Data Analysis and Generation of High-confidence Transcript Set—The reads were quality filtered for Phred-based base quality ($Q > 20$) using FastX tools. 99% of the reads passed the quality threshold and were used in downstream analysis steps. TopHat (version 1.4.1) with default parameters was used to align the reads against the Zv9 zebrafish genome assembly (14). Transcript assembly was carried out using Cufflinks (version 2.0). The RABT (Reference Annotation Based Transcript Assembly) option was used. An Ensembl transcript coordinate file (.gtf) was provided as a reference assembly file. Transcripts were assembled separately for each organ and were combined using Cuffcompare. Transcripts were also categorized (class codes) into known isoforms, novel isoforms, and intergenic transcripts by Cuffcompare (15). From the combined set of transcripts, a high-confidence set of transcripts was generated by filtering as shown in supplemental Fig. S1. Briefly, all the transcripts were filtered for fragments per kilobase of exon per million fragments mapped (FPKM) ≥ 1. From the remaining set, transcripts with Cufflinks class codes e, p, c, o, and s were eliminated. From transcripts with class codes u, i, x, and o (multi-exonic), transcripts smaller than 250 bp were eliminated. All transcripts of class codes = and j were retained. Transcripts for which peptide evidence was obtained were retained regardless of their class code and size. Protein coding potential was predicted for these high-confidence sets of transcripts using CPAT (16). Transcripts that had a coding probability greater than 0.38 were considered as potentially protein coding transcripts.

LC-MS/MS Analysis—Different organs (eye, brain, liver, spleen, intestine/pancreas, ovary, testis, muscle, heart, and head) were collected from ~100 adult fish (SAT strain). Zebrafish embryos were collected at 48 and 120 h post-fertilization. The samples were lysed in 2% SDS lysis buffer and in 8 M urea lysis buffer. Lysates were homogenized and sonicated, and protein estimation followed. Proteins from SDS lysates were separated on SDS-PAGE, and in-gel digestion was carried out as described previously (17). Briefly, the protein bands were destained, reduced and alkylated, and subjected

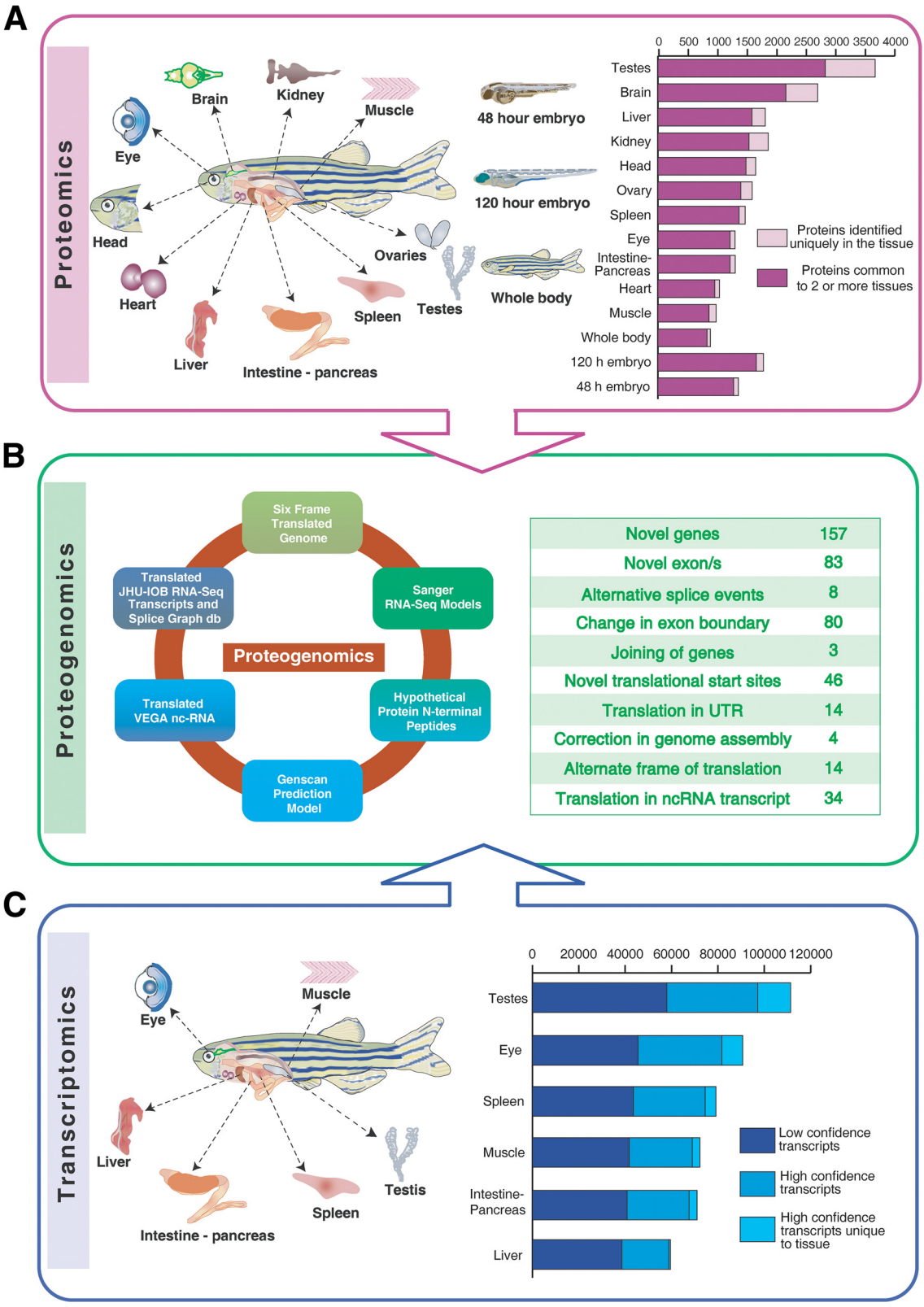

FIG. 1. **Integration of transcriptomic and proteomic data for improving genome annotation.** *A*, adult zebrafish organs and developmental stages used in proteomic analysis and total proteins and sample-specific proteins identified in the study. *B*, seven alternative databases used in proteogenomic analysis and summary of proteogenomics findings. *C*, adult zebrafish organs used in transcriptomic analysis, total transcripts, high-confidence transcripts, and sample-specific transcripts identified in each organ.

to in-gel digestion using trypsin and Lys-C at an 8:1 ratio. The peptides were extracted, vacuum dried, and stored at −80 °C until further analysis.

1-mg protein samples from urea lysates were used for in-solution trypsin digestion. Samples were reduced, alkylated, and digested using trypsin and Lys-C (8:1) overnight at 37 °C. The peptide digests were then desalted using a C_{18} cartridge and lyophilized. The lyophilized samples were reconstituted in basic reverse-phase liquid chromatography solvent A (10 mM tetraethyl ammonium bicarbonate (TEABC), pH 8.5), loaded on an XBridge C_{18} 5 μm 250 × 4.6 mm column (Waters, Milford, MA), and eluted with 0% to 100% solvent B (10 mM TEABC in acetonitrile, pH 8.5) with a 50-min gradient. The fractions collected were vacuum dried and then pooled by concatenation into 24 fractions.

Enrichment of N-terminal acetylated peptides was carried out using a slightly modified protocol described by Taouatas *et al.* (18). Briefly, purified peptides from in-solution digestion of protein extracts from four organ pairs were fractionated on a polysulfoethyl A column (PolyLC, Columbia, MD; 200 × 2.1 5 μm, 200 Å) using a low-ionic-strength buffer system (solvent A, 5 mM KH_2PO_4, 30% acetonitrile at pH 2.7; solvent B, 350 mM KCl, 5 mM KH_2PO_4, 30% acetonitrile). The first 15 fractions that were collected were pooled and re-fractionated into 24 fractions using basic reverse-phase liquid chromatography.

Peptide samples from in-gel digestion and basic reverse-phase liquid chromatography fractionation were analyzed on an Accurate Mass Q-TOF 6540 mass spectrometer interfaced with an HPLC Chip system (Agilent Technologies, Santa Clara, CA.). The samples were reconstituted in solvent A (0.1% formic acid) and loaded onto the HPLC chip trap column using an Agilent 1200 series capillary liquid chromatography system. Both trap and analytical columns embedded in the HPLC chip were made up of Zorbax 300SB-C_{18} with a 5-μm particle size. The peptides were eluted using a gradient of 5% to 40% solvent B (0.1% formic acid in 90% acetonitrile) over 50 min. Q-TOF was operated with a capillary voltage of 1800 V, a fragmentor voltage of 175 V, a medium isolation width of 4 *m/z*, and an energy slope of 3 V plus a 2-V offset. MS data were acquired using MassHunter data acquisition software (Version B.04.00, Agilent Technologies). MS spectra were acquired in the range of *m/z* 350–1,800, and this was followed by five MS/MS analyses with a scan range of *m/z* 50–2,000. The duty cycle was set to 2.1 s with one MS scan per second followed by five MS/MS scans per second. The precursor selection was based on preference to charge state in the order of 2+, 3+, and >3+ ions and a second level preference to abundance. Additionally, in-gel digested samples from zebrafish testis and spleen were analyzed on an Orbitrap Velos mass spectrometer. Enriched N-terminally modified peptide samples were analyzed on an Orbitrap Elite mass spectrometer. Both MS and MS/MS spectra were acquired in the Orbitrap analyzer at 60K and 15K resolution settings, respectively. Fragmentation was carried out in higher-energy collisional dissociation mode.

Proteomic Data Analysis—MS/MS spectral data were processed to generate Mascot generic format files using MassHunter (B.04.00) or Proteome Discoverer 1.3. The data were searched against a protein database from Ensembl-HAVANA annotation (release 70) combined with common contaminants like trypsin, keratin, and BSA (42,200 total sequences). The data were analyzed using Proteome Discoverer 1.3 (Thermo Scientific, Bremen, Germany) using Sequest (SCM build 59) and Mascot (Version 2.2) search algorithms. The parameters used for data analysis included trypsin as protease with up to one missed cleavage allowed. Carbamidomethylation of cysteine was specified as a fixed modification, and oxidation of methionine was specified as a variable modification. The minimum peptide length was specified as six amino acids. The mass error of parent ions was set to 20 ppm, whereas for fragment ions it was set to 0.05 Da. LC-MS/MS data were searched against a reversed-sequence database to calculate a 1% false discovery rate (FDR) threshold score. The FDR at each PSM score was calculated as - % FDR = (number of hits in reverse database at or above the score/total number of hits in target and reverse database at or above the score) × 100 (19). A parsimonious protein list was generated from the peptide list by grouping proteins in Proteome Discoverer. For quantitative analysis, intensity- and PSM-number-based expression values for each identified gene were calculated on similar lines of intensity-based absolute quantification values (20). The sum of the intensities of all the PSMs belonging to one gene was divided by the number of possible unique tryptic peptides from the gene. This value was normalized across experiments by dividing it with the total number of spectra acquired in the experiment (*e.g.* the total number of spectra from brain in-gel fractions). Finally, the ratio was log2 transformed. Gene functional classification was done using the Web-based DAVID resource, and significantly enriched gene sets ($p < 0.5$) were selected (21).

Identification of Novel Peptides Using Alternative Database Searches—Seven alternative databases were used for identifying novel peptides for proteogenomic analysis of the zebrafish genome. The seven databases used were (i) a six-frame translated genome database, (ii) a translated RNA-Seq transcript database (from this study), (iii) a translated Ensembl RNA-Seq transcript database, (iv) a splice graph database generated from RNA-Seq reads split across splice junctions, (v) *ab initio* prediction models from GENSCAN, (vi) a hypothetical N-terminal peptide database, and (vii) a three-frame translation of non-coding RNAs from VEGA annotation. Common contaminant sequences were added to each database prior to searching. A decoy database was created for each database by reversing the sequences in target databases. Peptide identification was carried out using X!Tandem (version CYCLONE, 2011.12.01) unless otherwise specified. Search parameters common to all searches were (a) an allowed precursor mass error of 20 ppm, (b) an allowed fragment mass error of 0.05 Da, (c) carbamidomethylation of cysteine was as a fixed modification, (d) oxidation of methionine as a variable modification, and (e) consideration only of tryptic peptides with up to one missed cleavage. All the novel peptides identified from each alternate database search were subjected to manual validation apart from filtering for 1% FDR. Specific details about the generation of each database, search parameters, and post-processing of the search outcomes are described below.

A six-frame translated genome database was created using the Zv9 version of the genome sequence downloaded from the Ensembl FTP server. Similarly, translated transcriptome databases were created using assembled transcripts from JHU-IOB and Sanger RNA-Seq data. Sanger RNA-Seq build was downloaded using the pearl API from the Ensembl "other features" database. (Transcripts falling in categories =, j, o, and c were translated in three frames; other transcripts were translated in six frames.) The protein databases thus created consisted of stop-codon-to-stop-codon translation of the template sequence. Sequences that were shorter than seven amino acids in length were not included. A splice graph database was generated from RNA-Seq read alignments. A splice graph database is a non-redundant compact database of splice junction peptide sequences derived from RNA-Seq reads split across introns. The database is created by generating a graph in which genomic intervals (exonic regions) correspond to nodes, and edges correspond to pairs of exons that are putatively spliced together. A detailed method of splice graph creation and conversion to an MS search compatible FASTA database can be found in Ref. 22. The splice graph database and the six-frame translated genome database were searched using the MS-GFDB (version 20120106) search algorithm. Search results from the MS-GFDB algorithm were filtered for 1% FDR calculated as explained in Ref. 22. Spectra that matched to multiple sequences with

equal scores were not considered for further analysis. Prediction models from the *ab initio* gene prediction algorithm GENSCAN were downloaded from the Ensembl FTP, and VEGA non-coding RNA sequences for processed pseudogenes, processed transcripts, pseudogenes, transcribed unprocessed pseudogenes, and unprocessed pseudogenes were downloaded using Ensembl BioMart; the sequences were translated in three frames. Additional variable modification of protein N-terminal acetylation was used for searching the Genscan prediction database.

The hypothetical N-terminal tryptic peptide database was created by fetching all the peptide sequences that began with methionine and ended with K/R from the six-frame translated genome database. Peptide sequences with up to one missed cleavage and lengths ranging from 6 to 25 amino acids were considered. Sequest and X!Tandem were used for peptide identification. Additional variable modification of peptide N-terminal acetylation was specified for these searches. For X!Tandem searches, the database without subpeptides was used as the "quick acetyl" option available in X!Tandem searches.

Workflow for Manual Genome Annotation—Peptide sequences identified from the alternate database searches were filtered for 1% FDR and compared with the protein database (Ensembl Genebuild 70) to identify novel peptides. These novel peptide PSMs were further checked via manual inspection for validity of the peptide identification. The major criteria considered for manual evaluation included (a) assignment of all intense peaks (intense unassigned peaks were checked to see whether they were arising from internal fragment ions); (b) identification of the majority of the y series of ions; (c) low-*m/z*-range b ions, that is, b_1, b_2, and b_3 ions and a_2 and a_4 ions also observed in a typical spectrum; (d) whether an immonium ion indicated the presence of an amino acid that was not present in the assigned peptide sequence (if so, the PSM was rejected); (e) whether a Y_1 ion was present that confirmed a peptide ending either with K (*m/z* 147.11) or with R (*m/z* 175.12); (f) whether any un-assigned fragment was present, especially from the higher *m/z* range, that indicated the presence of an amino acid that was not a part of an assigned sequence (if so, the PSM was rejected); (g) whether missed cleavages were followed by acidic amino acids, that is, E and D; (h) whether many assigned peaks were from the noise level (if so, the PSM was rejected); and (i) whether neutral loss ions were observed for peptides containing methionine ($SOCH_4$), glutamine/asparagine (NH_3), and glutamic acid/aspartic acid (H_2O). Integrative Genomics Viewer (IGV) was used as a visualization tool for manual genome annotation using these novel peptides (23). IGV version 2.2 was downloaded from the Broad Institute server and installed on a standalone Windows system. Novel peptides, Ensembl-HAVANA transcripts, JHU-IOB RNA-Seq models, Sanger RNA-Seq models, and gene prediction models from Genscan were tracked against the Zv9 assembly of the zebrafish genome on the IGV genome browser. For manual genome annotation, genomic regions where novel peptides and RNA-Seq models mapped were examined, and novel genes or changes in gene structures were determined.

For functional analysis of novel genes, the protein sequence was obtained from the longest ORF in the frame of the peptides. If the peptide matched to multiple transcripts, the transcript that had the longest ORF in the frame of the peptides was selected. Functional prediction was carried out by searching for conserved domains and known protein homologues using SMART and Blast2GO 2.6.0 (E-value threshold of 1E-03, HSP length of 35) (24).

RT-PCR—Total RNA was isolated from muscle, liver, intestine, pancreas, ovary, brain, testis, eye, spleen, kidney, heart, and 48-h embryo and converted to cDNA. RT-PCR validation was carried out for novel genes, genome assembly error, novel exons, and novel splice events. PCR primers were designed in the exonic regions of the alternate gene models. Wherever possible, primers were designed across introns to rule out genomic contamination. Specific amplicons were purified and sequenced. High-quality sequences were deposited in the NCBI EST database.

Data Availability in Public Repositories—Raw sequencing data (.bam files) from the transcriptomic profiling experiment have been deposited in the SRA database (www.ncbi.nlm.nih.gov/sra). They can be accessed using identifiers SRA060234 and SRX204106. Raw spectral data and search results from proteomic profiling have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository and can be accessed using the identifier PXD000479 (25).

RESULTS

Overview of Transcriptomic Analysis—High-throughput mRNA sequencing was carried out for six adult organs of zebrafish: eye, spleen, muscle, liver, testis, and intestine/pancreas. On average, 90 million reads were obtained per organ. We chose these tissues in order to complement transcriptomic data from ovary, whole body, and developmental stages that have already been integrated into genome annotations from Ensembl (2). Read alignment was carried out using TopHat, and transcript assembly was carried out using Cufflinks (14, 15). About 73% of the reads were aligned to the genome (Zv9). About 30,000 to 50,000 transcripts were identified in each tissue, and a total of 147,104 unique transcripts were identified from all tissues (derived from 78,104 unique loci). We filtered these transcripts as shown in supplemental Fig. S1 to obtain a set of 69,206 high-confidence transcripts. 46,864 of these transcripts were found to be novel when compared with the Ensembl gene set (Fig. 2*A*). Transcript products were identifiable for 23,915 genes of the 32,469 annotated Ensembl genes (Fig. 2*B*). The protein-coding potential of these transcripts was evaluated using CPAT (16). 72% of these (*i.e.* 49,820 transcripts) were predicted to be protein-coding. When compared with Ensembl protein-coding transcripts, 29,790 potentially protein-coding transcripts were uniquely identified in this study (Fig. 2*C*).

The overall expression level of known transcripts was higher than that of novel isoforms and novel intergenic transcripts. Further, transcripts from protein-coding genes were more abundant than non-coding transcripts (Fig. 2*D*). Although the current Ensembl genebuild was improved by incorporating RNA-Seq data described by Collins *et al.*, the RNA-Seq data were not resolved at the transcript level (2). We used Cufflinks, which generates potential multiple transcripts of a gene, to identify novel transcriptional isoforms for 22,585 Ensembl genes (transcripts with class code j). Transcripts from intergenic regions (u category) indicate potential novel genes. We identified 11,342 such transcripts belonging to 9,404 loci. About one-third of these novel intergenic transcripts are expected to be protein-coding as per CPAT prediction, and the remainder are possibly long non-coding RNAs.

Careful analysis of proteome/transcriptome profiles can provide insights into the biology of tissues. In our analysis,

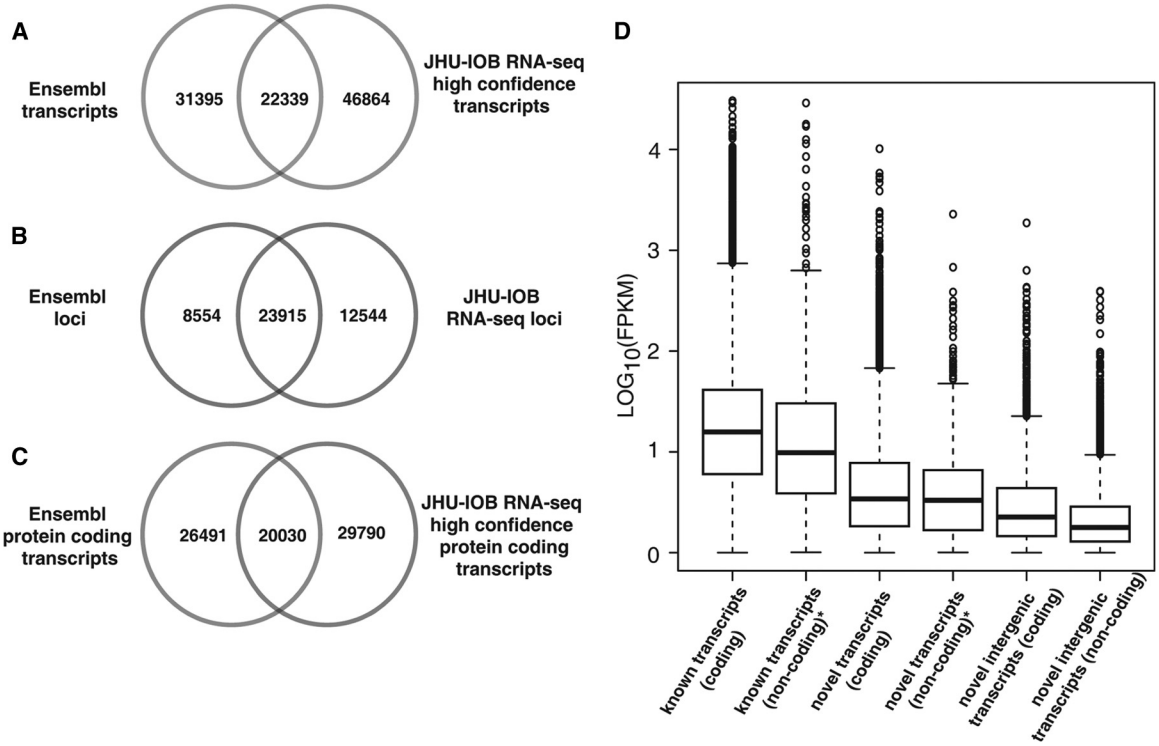

FIG. 2. **Identification of novel transcripts and novel protein coding loci by transcriptomic profile.** *A*, overlap between Ensembl transcripts and high-confidence transcripts identified in this study. *B*, overlap between Ensembl genes and RNA-Seq loci identified in this study. *C*, overlap between Ensembl protein coding loci and the novel RNA-Seq transcripts predicted to be protein coding by CPAT. *D*, distribution of FPKM values in known transcripts (class code =), novel transcripts (class code j), and novel intergenic transcripts (class code u). The coding and non-coding sets were generated using CPAT predictions. *Transcripts of miRNA and very short genes were excluded because of their very high FPKM values due to their short length.

liver showed the least diversity in terms of the number of transcripts or transcribed loci. Remarkably, ~20 million reads out of 105 million reads from liver mapped to seven genes belonging to the vitellogenin gene family clustered on chromosome 22. This was expected because vitellogenin genes are highly expressed in the liver (26), and it correlates with the proteomic data in which the five most abundant proteins based on spectral counts were indeed encoded by vitellogenin genes. Testis showed the maximum number of transcripts, as well as the highest number of known and novel loci (Fig. 1, bottom panel). About 26% of the transcripts from testes were predicted to be non-coding, in contrast to the average 16% of transcripts that were predicted to be non-coding in other organs. This is consistent with the findings of Ulitsky *et al.*, who found the greatest number of testis-specific poly (A) sites using the 3P-seq technique; many of those sites were speculated to be derived from as yet unannotated gene models (27).

Overview of Proteomic Analysis—High-accuracy MS/MS data were acquired for 10 adult organs, adult head, whole adult body, and two developmental stages. An MS/MS database search using the Sequest and Mascot search engines led to the identification of 6,975 proteins based on 937,990 PSMs (46,273 unique peptide sequences) at a 1% FDR cutoff. Complete lists of peptides and proteins identified in this study are provided in supplemental Tables S1 and S2, respectively.

As shown in Fig. 3*A*, >76% of the peptide identifications were identified with <5 ppm mass error. Further, more than 30% of the proteins identified in this study had >30% sequence coverage (Fig. 3*B*). The histogram in the top panel of Fig. 1 shows the total number of proteins and the number of proteins uniquely identified in 10 organs. The greatest number of unique proteins was identified in testis, followed by eye and brain. From a total of 46,273 peptides identified from a protein database search, 12,279 were found to be exon–exon junction-spanning peptides. From these data, we were able to confirm splice junctions for 4,635 Ensembl annotated genes (supplemental Table S3). Protein N-terminal acetylation modification was included in Mascot searches carried out against the protein database. From this, we confirmed the annotation of translational start sites in 1,189 genes through the identification of 1,194 N-terminally acetylated peptides (Fig. 3*C*) (supplemental Table S4).

Genes designated as uniquely identified or highly expressed in a particular organ were identified from intensity and PSM values. Tissue-related functions such as involvement in sensory perception of light stimulus and protein kinase activity were found to be enriched in the eye. Similarly,

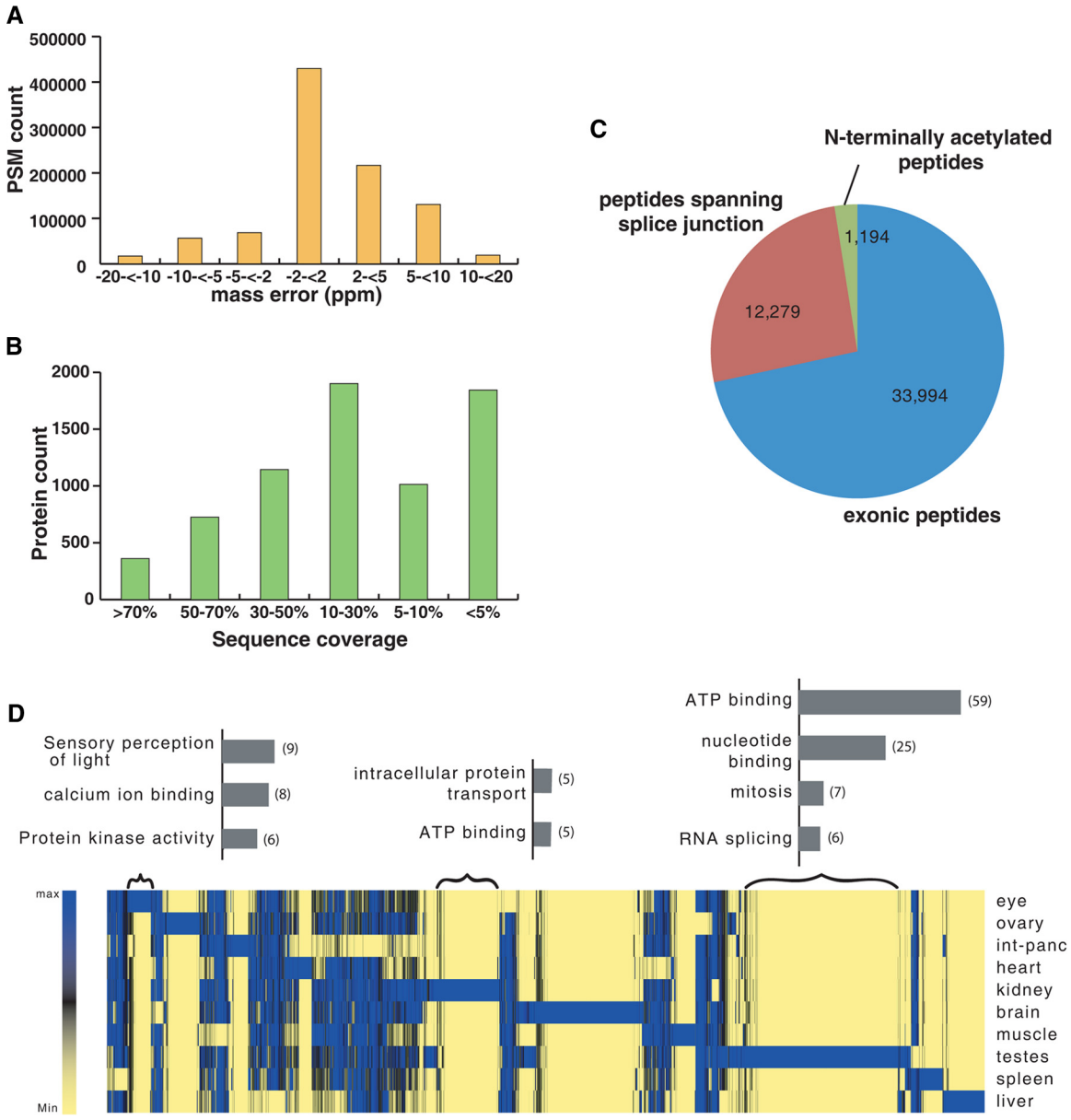

FIG. 3. **Proteomic profiling of zebrafish.** *A*, distribution of PSMs according to mass error. *B*, distribution of proteins according to the extent of sequence coverage. *C*, peptides used for confirmation of current annotations from the protein database search. *D*, expression analysis using intensity-based absolute quantification protein expression values. Functional categories enriched in different organ-specific gene sets were obtained using DAVID ($p < 0.5$). The figure shows enriched categories in eye, kidney, and testis gene sets. Numbers in parentheses indicate the number of genes in the functional category.

genes involved in RNA splicing, nucleotide binding, and mitosis were enriched in the testis, indicating elevated transcriptional activity and cell proliferation (Fig. 3*D*). Recently, a study reported proteomic profiling of nine zebrafish organs using SILAC-labeled standards (28). Comparison with the current study revealed many common observations such as high expression of ependymin in brain and ventricular myosin heavy chain-like in heart.

Refinement of Genome Annotation via Integrative Proteogenomic Analysis—Proteomic and transcriptomic data were integrated for the proteogenomic analysis described in this study. We used both assembled transcripts and directly translated RNA-Seq data to identify peptides from MS/MS data. Novel peptides were identified through comparison with the protein sequences from Ensembl genebuild 70, which was the latest release at the time of analysis. We used translated transcript sequences in addition to translated genomes for peptide identification because the transcript sequences provide a compact sequence database resulting in a reduced search space and at the same time provide exon–exon spanning peptides, which are missing in the six-frame translation of the genome. This was evident in our analysis, where 95%

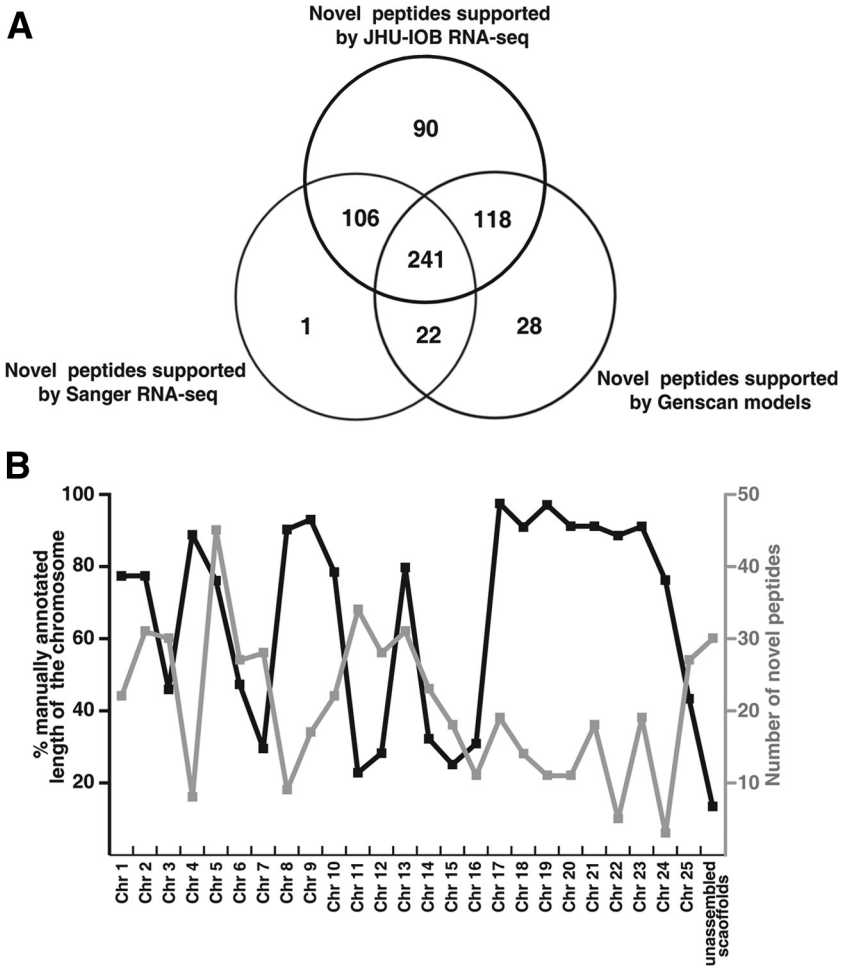

FIG. 4. **Novel peptides.** *A*, novel peptides exclusively supported by the novel RNA-Seq transcripts identified in this study along with Sanger RNA-Seq transcripts and Genscan models. *B*, novel peptides identified in 25 chromosomes and unassembled contigs plotted along with the percent length of chromosomes for which manual annotation was carried out.

of novel peptide annotations were supported by RNA-Seq data (JHU-IOB and Sanger RNA-Seq models) (Fig. 4*A*).

The distribution of novel peptides identified from different chromosomes is shown in Fig. 4*B*. As a general trend, smaller numbers of novel peptides were mapped to those chromosomes where manual genome annotation (Vega) was mostly completed. A high number of novel peptides were identified on chromosomes 3, 6, 7, 11, 12, and 25 and unassembled contigs where manual annotation seemed to be less complete. Surprisingly, however, a high number of novel peptides were also identified in chromosomes 2, 5, and 13, where manual annotation was completed for >70% of the chromosome. This highlights the benefit of proteogenomic analysis, which can complement other methods of genome annotation even for well-annotated sequences. Although the alternative gene and transcript models generated from transcript data alone could potentially be used to propose many changes in the genome annotation, in the manual annotation process adopted by our group, only those RNA-Seq-based revisions that were also supported by peptide evidence were considered. Findings from our proteogenomic analysis are discussed in the following sections (supplemental Tables S5*A*–S5*J*, summarized in Fig. 1*B*).

Identification of Gene Models Split in the Genome Because of Errors in the Genome Assembly—In four cases, we found that the gene sequence and RNA-Seq transcripts were split in the genome owing to errors in genome assembly. Identification of novel peptides mapping to intergenic regions led us to detect four cases of misassembled contigs (supplemental Table S5*A*). As shown in Fig. 5*A*, peptides MSHADSALLADIMDEAR and RQVGVVYSQDSQ mapped to an intergenic region on chromosome 18 corresponding to genome coordinates of 680,112–680,198. These peptides belong to the carboxy-terminal of the dihydrodiol dehydrogenase enzyme as identified by an analysis of orthologous proteins in other species. The remaining part of this protein aside from these two peptides (ENSDARG00000019081) is located 77 kb downstream on the same chromosome (genome coordinates 756,212 to 768,377). A partial RNA-Seq transcript is also present where the novel peptides were mapped; however, a closer look at the read alignment clearly showed that the mate pairs of these reads also aligned 77 kb downstream, where the amino-terminal part of the gene is located. Similarly, in the case of genes ENSDARG00000044097 (stomatin like 2) and ENSDARG00000018562 (coatomer subunit ϵ), partial gene sequences were found to map ~530 kb upstream and 140 kb

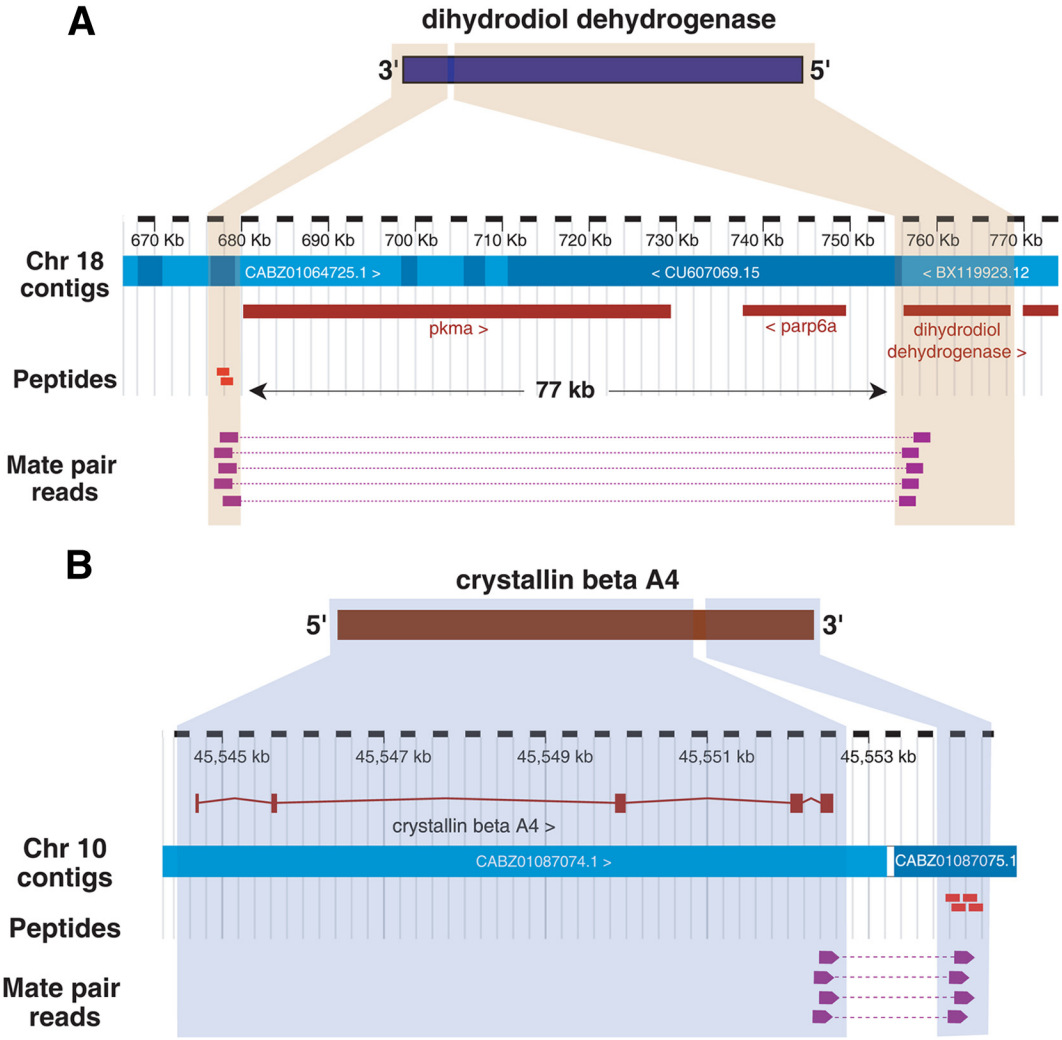

FIG. 5. **Discovery of genome assembly errors.** *A*, two novel peptides were found to map to an intergenic region on chromosome 18. Orthology analysis showed that they belong to the carboxyl terminal of dihydrodiol dehydrogenase protein. The remainder of the 5′ part of the gene is located 77 kb downstream on the same chromosome. Reads aligning to the peptide region revealed that the mate pairs of these reads also aligned to a region 77 kb downstream in the last exon of the annotated gene, providing a second line of evidence for error in genome assembly. *B*, four peptides were found to map immediately downstream of the gene crystalline β A4 on the opposite strand. Orthology analysis showed that these peptides belong to the carboxyl terminal of the crystallin β A4 protein, showing that the downstream contig was assembled in the opposite orientation. RNA-Seq read alignment also supported this observation where the mate pair reads map in the same orientation.

downstream, respectively, again owing to genome assembly errors (supplemental Figs. S2*A* and S2*B*). Peptides and RNA-Seq reads both supported these assertions. In the case of the third example (coatomer subunit ε), we not only found the missing amino terminus of the protein but also confirmed its translation initiation site based on an N-terminally acetylated peptide. The last case involved the crystallin β A4 gene (ENSDARG00000024548), where peptides from the missing carboxy-terminal sequence mapped immediately downstream but on the opposite strand, suggesting that this genome contig was incorrectly assembled in the reverse orientation (Fig. 5*B*). This observation is also supported by unusual read alignments in which the mate pair reads map on the same strand, supporting the conclusion that adjacent contigs in this region are assembled in the wrong orientation. All of these examples of genome assembly error were validated using RT-PCR.

Identification of Novel Protein-coding Genes—Novel peptides identified from MS/MS searches against translated genome, translated transcriptome, the splice graph database, and Genscan prediction models were further investigated for the identification of novel genes, novel exons, or changes in exon coordinates. Transcript models from the RNA-Seq/prediction gene set that were supported by novel peptides and those that did not overlap with any annotated exonic regions were considered to be novel gene products. A total of 157 novel genes were identified based on this analysis, of which 49 were based on evidence from two or more peptides. 87

novel genes had RNA-Seq models from both Sanger and JHU-IOB datasets, and 56 novel genes were exclusively supported by JHU-IOB RNA-Seq data. 12 novel genes had only peptide-level evidence. These findings underscore the importance of carrying out an integrated multi-omics analysis for genome annotation. We inferred functions for novel genes based on sequence similarity and protein domain using Blast2GO and SMART domain analysis tools, respectively (24). For 89 of 157 novel genes, the domain structure, functional category, or both could be assigned. Some of the important findings of this analysis were novel genes coding for von Willebrand factor (novel gene 18), filamin (novel gene 4), spectrin (novel gene 47), acetyltransferase (novel gene 105), and vinculin (novel gene 12)-like proteins. Supplemental Table S5*B* lists the novel genes found in this study with the corroborating peptide evidence, transcriptome evidence, protein domain, and functional categorization.

We validated 28 randomly selected novel genes at the mRNA level using RT-PCR. Nine of the validated examples were supported by a single peptide. This demonstrates that in proteomic studies, single-peptide identifications could be just as valid as multi-peptide identifications and could be corroborated through additional analysis similar to what was performed in this study. Notably, 23 of these novel genes were incorporated in a later version of Ensembl (Genebuild 75). An example of a novel gene proposed using nine unique intergenic peptides is shown in Fig. 6. RNA-Seq transcript data suggested an 11-exon gene model, which is in agreement with several of the identified peptides. A domain analysis of the 350-amino-acid-residue-long protein product of this novel gene suggests that it is an acetyltransferase.

Identification of Novel Exons and Their Effect on Domain Architecture of Proteins—Novel exons were identified for 83 annotated gene models. In 23 cases, exons were added to intronic regions (*i.e.* missed internal exons), and in the remainder they were added to 3′ or 5′ termini extending the genomic coordinates of the gene. In 31 genes, multiple exons were added based on proteomic and transcriptomic evidence. Upon further analysis, domain structures of many of these proteins were found to be altered, or significant additions were made to the overall architecture. For example, a phosphotransferase domain and two ankyrin domains were added to the product of the TEX14 gene (ENSDARG00000090187), which was extended by more than 500 amino acids. In another case, three RNA recognition motifs were added to the gene encoding heterogeneous nuclear ribonucleoprotein R (ENSDARG00000014569). Some other examples included addition of a phosphofructokinase domain, a transmembrane domain and peptidase dimerization domains added to phosphofructokinase (ENSDARG00000042375), damage-specific DNA binding protein (ENSDARG00000074431), and carnosine dipeptidase (ENSDARG00000069583) proteins. In the case of the ENSDARG00000093267 gene, which codes for a protein homologous to potassium channel SKOR protein, novel exons added on the basis of both proteomic and transcriptomic evidence formed a link between two different transcripts of this gene (ENSDART00000137234 and ENSDART00000145759) that were designated as containing incomplete coding sequences by Ensembl. The product of the merged coding sequences was found to have an additional cyclic nucleotide-monophosphate binding domain. Joining of these transcripts was validated by RT-PCR, along with two more examples of novel exons. A list of genes with a novel exon (or exons) added is provided in supplemental Table S5*C*.

One of the commonly observed errors in genome annotation is a single transcriptional unit being annotated as two different genes. For three such instances, we found both proteomic and transcriptomic evidence, which corrected this annotation error (supplemental Table S5*D*). Fig. 7*A* illustrates one such example where seven peptides were found to map to an intergenic region between Ensembl genes that code for partial dynein heavy peptides (ENSDARG00000004221 and ENSDARG00000011633). The RNA-Seq model showed the presence of nine exons that formed a bridge between these two genes, arguing for the presence of a single gene. The merged model was included in the subsequent Ensembl release genebuild 73 but removed from genebuild 75. The longer merged protein product shows the presence of an additional well-conserved domain, DHC_N2, which was missing from the annotated gene. Two of these examples were validated by RT-PCR.

Changes in the Annotation of Exon Boundaries and Identification of Novel Splice Events—Based on peptide evidence, we cataloged changes in exon coordinates in 80 genes (supplemental Table S5*E*). Four categories of novel peptides led to the identification of these examples: (i) intergenic peptides, (ii) intronic peptides, (iii) peptides partially mapping to exons, and (iv) peptides mapping to exon junctions arising from alternate donor and acceptor site usage (identified from RNA-Seq, splice graph, and Genscan database searches). Some of the prominent genes in which a change in exon boundaries was proposed were clathrin light chain B, DEAD box polypeptide 6, mitofusin 1, SERPINE1 mRNA binding protein 1, eukaryotic translation initiation factor 2B, subunit 1 α, and complement component 3. Apart from these, we identified eight exon skipping events by searching proteomic data against translated RNA-Seq transcripts, the splice graph database and Genscan predicted transcripts. Genes in which this alternate splicing was observed include spectrin α 2, obscurin, neurofascin homolog a, autophagy related 7, complement component 3, solute carrier family 35A2, huntingtin interacting protein 1 related, and coiled-coil domain containing 15. Two of these splice events were validated by RT-PCR and sequencing (supplemental Table S5*F*).

Correction of Gene Structures Using Exonic Peptides—There are nine glutathione peroxidase coding genes predicted in zebrafish genome: gpx1a, gpx1b, gpx2 (si:ch73–199k24.2), gpx3, gpx4a, gpx4b, gpx7, gpx8, and si:ch73–111m19.2. We

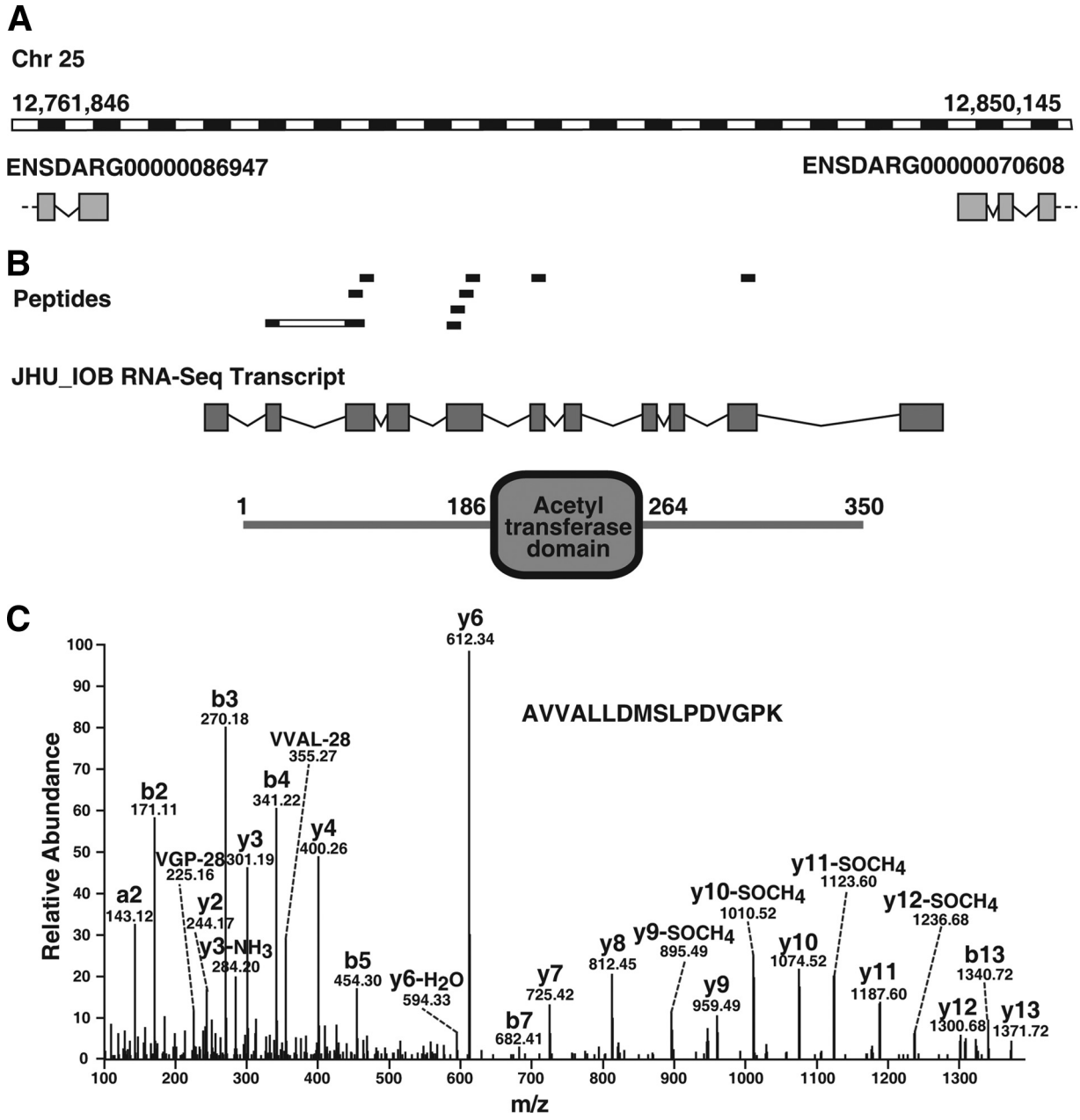

FIG. 6. **Identification of a novel gene.** *A*, nine peptides mapped to intergenic region on chromosome 25. The RNA-Seq model suggests the presence of a novel gene in this region. *B*, Pfam domain structure of the novel gene indicating acetyltransferase function. *C*, labeled MS/MS spectrum of a representative peptide AVVALLDMSLPDVGPK.

identified novel peptides that mapped to the UTRs of three of these genes: gpx1a, gpx4a, and gpx4b. Glutathione peroxidase is known to possess a selenocysteine in the active site. Peptides mapping to the UTR revealed a systematic annotation error in which this key selenocysteine amino acid was missing from the annotated sequences in zebrafish (Fig. 7*B*). Because of this, the translation initiation codons for these three genes were annotated downstream of the codon for selenocysteine, UGA, a non-sense codon. When we looked at the remaining six glutathione peroxidase coding genes, we observed that proteins encoded by two of these genes, gpx7 and si:ch73–111m19.2, lacked the initiator methionine, whereas that encoded by gpx1b lacked the selenocysteine residue (with the annotated start site located downstream of the selenocysteine codon), clearly indicating the need to revise these annotations.

Three genes that code for the thymosin family of proteins are annotated in the zebrafish genome. We found that peptides mapped to the 5′ UTR of the ENSDARG00000093886 gene, suggesting that there might be an error in the frame of

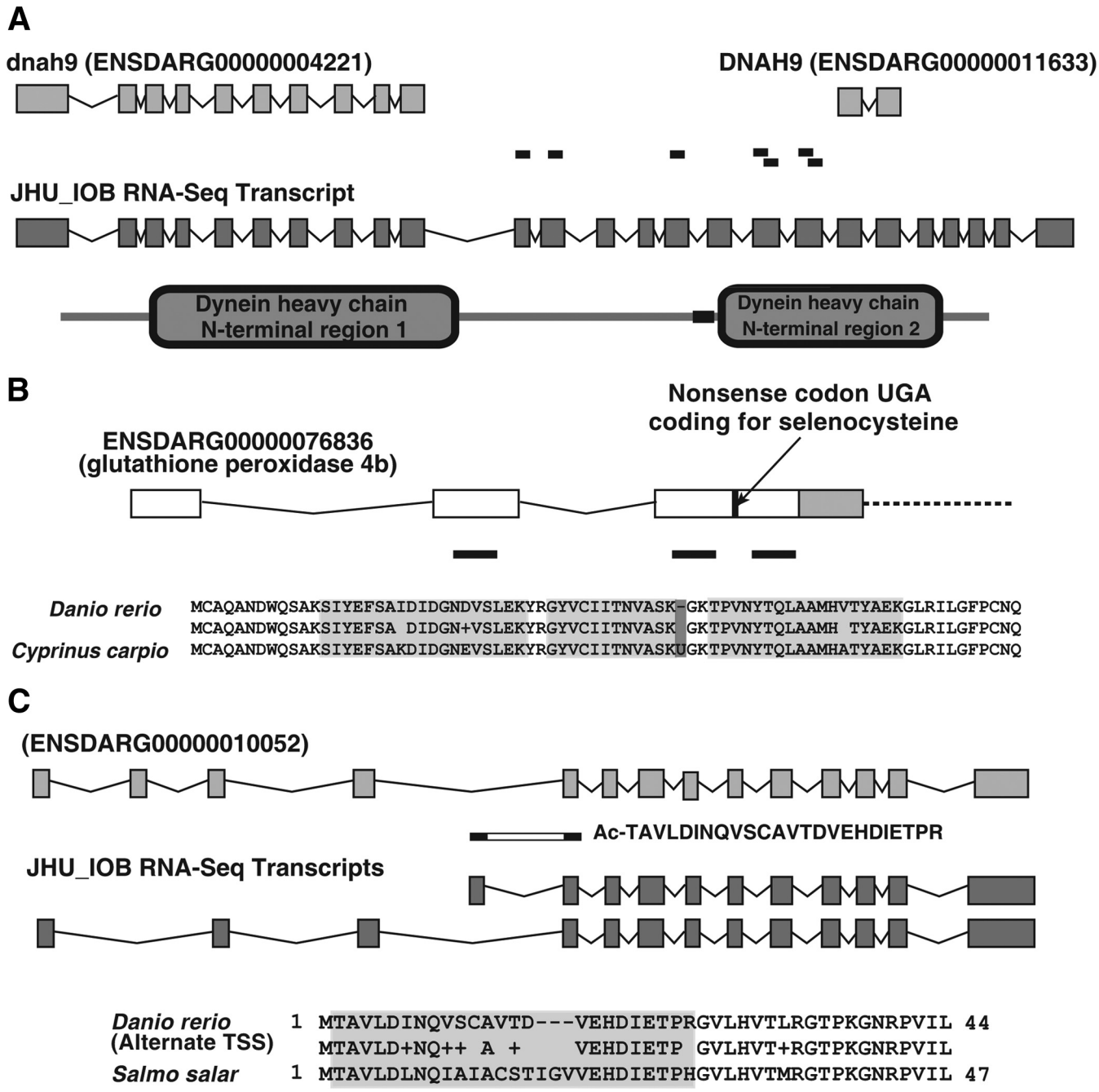

FIG. 7. *A*, merging of two genes. Seven peptides mapping to the intergenic region between genes ENSDARG00000004221 and ENSDARG00000011633 and an RNA-Seq transcript model suggest a single transcriptional unit. An additional Dynein heavy chain N-terminal domain is present in the merged transcript. *B*, correction in the amino-terminal annotation of glutathione peroxidase gene. Three peptides mapping to the 5′ UTR of the glutathione peroxidase 4b gene (ENSDARG00000076836) indicate the presence of a protein-coding region upstream of the annotated translational start site. N-terminal extension of the coding sequences by correcting a non-sense codon (UGA) to a selenocysteine codon is supported by orthologous protein from carp. *C*, detection of an alternate translational start site by N-acetylated peptide. An acetylated peptide mapping partially in the intron of ENSDARG00000010052 indicates the presence of an alternate translational start site. A shorter transcript from RNA-Seq data supports the alternative translational start site downstream of the annotated translational start site. The alternative translational start site is supported by orthologous protein in *Salmo salar.*

translation. The protein sequence derived from this corrected ORF revealed that the protein belongs to the thymosin family, making the ENSDARG00000093886 gene the fourth thymosin family gene in the zebrafish. 13 more examples of peptides in UTRs (six in 5′UTR and seven in 3′UTR) were identified extending their coding sequences (supplemental Table S5*G*). In two cases where peptides mapped to the 3′ UTR (genes ENSDARG00000019406 and ENSDARG00000018351), the annotated protein sequences did not terminate in a stop codon, enabling correction of the 3′ termini of the corresponding coding sequences. (Annotation of ENSDARG00000019406 was corrected in the next version of Ensembl (Genebuild 73).) Apart from thymosin genes, 13 additional examples of alternative frames of translation were identified in this study (supplemental Table S5*H*).

Peptides identified from the MS/MS searches against Vega and other ncRNA annotations were investigated in greater detail to identify protein-coding genes that were incorrectly annotated as non-coding RNAs. We found 34 cases of translated ncRNA transcripts (supplemental Table S5*J*), of which 31 were labeled as non-coding RNA genes categorized as "processed transcript" or "retained intron transcripts," and the remaining 3 were designated as pseudogenes (1 processed pseudogene and 2 unprocessed pseudogenes).

Identification of Alternate Translation Initiation Sites—In genome annotation efforts, identification of N-terminal acetylated peptides can both confirm and correct translational start sites. Many transcripts are known to have AUG codons upstream of annotated translational start sites. Generally, translation initiation sites are assigned based on the presence of a consensus Kozak sequence (CCACCaugG) or a sequence that is very close to it. However, we have shown that in a large proportion of transcripts, the sequence around the presumed initiator codon is often less conserved than what is generally believed (29). Given this limitation of sequence-based methods, the use of proteomic data-derived experimental evidence could provide valuable information for the determination of true translation initiation sites.

Protein N-terminal peptides are considered as tryptic peptides by search algorithms; however, in three- or six-frame translated nucleotide sequence databases, N-terminal peptides appear semi-tryptic. Allowing the identification of semi-tryptic peptides in MS/MS searches vastly increases the search space, along with the associated false discovery rate. To circumvent this, we created a database of potential N-terminal peptides beginning with M and ending in arginine or lysine derived from the six-frame translation of the genomic sequence. N-terminally acetylated peptides identified by searching against this database and against Genscan models were compared with known protein annotations to designate start sites as novel. Using this approach, we identified 46 instances of alternate translation initiation sites (supplemental Table S5*I*). In 10 of these cases, the alternate translation start site mapped upstream of the known start site, whereas in 37 examples the novel start site mapped downstream of the annotated site. Notably, in 16 out of 46 of these cases, the annotated protein sequence lacked an initiator methionine. In 19 cases, we found orthologous protein evidence supporting the novel translational start site, and for eight genes we found peptide evidence for both annotated and alternate translational start sites.

Important genes for which novel translational start sites were identified included hemoglobin α adult-1, symplekin, importin 13, von Hippel-Lindau binding protein 1, solute carrier family 6 - member 3, myosin heavy polypeptide 2, ubiquitin carboxyl-terminal esterase L1, glycyl-tRNA synthetase, and RNA polymerase II polypeptide D. One example of an alternate translation start site is illustrated in Fig. 7*C*; we identified an N-terminal acetylated peptide derived from the fourth intron of the NDRG3b gene (ENSDARG00000010052). Importantly, a shorter transcript isoform from the RNA-Seq data, which included a novel exon located in the fourth intron, confirmed this alternate start site annotation. Finally, we also identified N-terminal acetylated peptides confirming translational start sites in 22 of the 157 novel genes reported in this study.

Discussion and Conclusions—Here, we describe an annotation method integrating transcriptomic and proteomic data that has not been previously explored for annotation of the zebrafish genome. In the past, transcriptomic profiling of zebrafish was reported for the whole fish body as one sample and for different embryonic stages (2, 30). We chose six adult tissues for which transcriptomic profiling had not been previously performed with the goal of capturing tissue-specific transcriptome information. In addition to novel protein-coding genes, we identified many potential lncRNA transcripts/genes in this study. However, we did not pursue validation and functional analysis for these novel non-coding transcripts, as our focus was mainly on the curation of protein-coding genes. Nevertheless, our analysis opens up the possibility of carrying out more detailed investigations along the lines of a recent study on embryonic lncRNAs (30). One unique feature of our study was the additional manual annotation of all novel events, including manual validation of peptide spectrum matches. Accurate annotation of all genomes is obviously highly desirable. However, this is especially true for the genomes of model organisms, as these frequently serve as templates for the annotation of newer genomes. Our findings are one step forward in achieving accurate annotation for zebrafish and should serve as a template for similar efforts in the future.

Acknowledgments—We thank the Department of Biotechnology (DBT) of the Government of India for research support to the Institute of Bioinformatics, Bangalore. S.M.S., K.K.D., S.R., G.D., and N.S. are recipients of Research Fellowships from the University Grants Commission (UGC), India. B.M., S.M.P., S.B.D., V.N., and J.S. are recipients of Senior Research Fellowships from the Council for Scientific

and Industrial Research (CSIR), India. H.G. is a recipient of an early career fellowship from the Wellcome Trust–DBT India Alliance.

S This article contains supplemental material.

c To whom correspondence should be addressed: Akhilesh Pandey, M.D., Ph.D., McKusick-Nathans Institute of Genetic Medicine and Departments of Biological Chemistry, Oncology and Pathology, 733 N. Broadway, 527 BRB, Baltimore, MD 21205, E-mail: pandey@jhmi.edu; Steven D. Leach, M.D., McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, 733 N. Broadway, 471 BRB Baltimore, MD 21205, E-mail: sleach1@jhmi.edu; Charles Wang, M. D., Ph.D., M.P.H., Center for Genomics and Division of Microbiology & Molecular Genetics, School of Medicine, Loma Linda University, 11021 Campus Street, Loma Linda, CA 92350, E-mail: chwang@llu.edu.

REFERENCES

1. Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G. J., White, S., Chow, W., Kilian, B., Quintais, L. T., Guerra-Assuncao, J. A., Zhou, Y., Gu, Y., Yen, J., Vogel, J. H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S. F., Laird, G. K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Eliott, D., Threadgold, G., Harden, G., Ware, D., Mortimer, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C. M., Ersan-Urun, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberlander, M., Rudolph-Geiger, S., Teucke, M., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Carter, N. P., Harrow, J., Ning, Z., Herrero, J., Searle, S. M., Enright, A., Geisler, R., Plasterk, R. H., Lee, C., Westerfield, M., de Jong, P. J., Zon, L. I., Postlethwait, J. H., Nusslein-Volhard, C., Hubbard, T. J., Roest Crollius, H., Rogers, J., and Stemple, D. L. (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496,** 498–503
2. Collins, J. E., White, S., Searle, S. M., and Stemple, D. L. (2012) Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res.* **22,** 2067–2078
3. Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenas, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010) Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6,** 450
4. Evans, V. C., Barker, G., Heesom, K. J., Fan, J., Bessant, C., and Matthews, D. A. (2012) De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Methods* **9,** 1207–1211
5. Peterson, E. S., McCue, L. A., Schrimpe-Rutledge, A. C., Jensen, J. L., Walker, H., Kobold, M. A., Webb, S. R., Payne, S. H., Ansong, C., Adkins, J. N., Cannon, W. R., and Webb-Robertson, B. J. (2012) VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data. *BMC Genomics* **13,** 131
6. Mohien, C. U., Colquhoun, D. R., Mathias, D. K., Gibbons, J. G., Armistead, J. S., Rodriguez, M. C., Rodriguez, M. H., Edwards, N. J., Hartler, J., Thallinger, G. G., Graham, D. R., Martinez-Barnetche, J., Rokas, A., and Dinglasan, R. R. (2013) A bioinformatics approach for integrated transcriptomic and proteomic comparative analyses of model and non-sequenced anopheline vectors of human malaria parasites. *Mol. Cell. Proteomics* **12,** 120–131
7. Chaerkady, R., Kelkar, D. S., Muthusamy, B., Kandasamy, K., Dwivedi, S. B., Sahasrabuddhe, N. A., Kim, M. S., Renuse, S., Pinto, S. M., Sharma, R., Pawar, H., Sekhar, N. R., Mohanty, A. K., Getnet, D., Yang, Y., Zhong, J., Dash, A. P., MacCallum, R. M., Delanghe, B., Mlambo, G., Kumar, A., Keshava Prasad, T. S., Okulate, M., Kumar, N., and Pandey, A. (2011) A proteogenomic analysis of Anopheles gambiae using high-resolution Fourier transform mass spectrometry. *Genome Res.* **21,** 1872–1881
8. Prasad, T. S., Harsha, H. C., Keerthikumar, S., Sekhar, N. R., Selvan, L. D., Kumar, P., Pinto, S. M., Muthusamy, B., Subbannayya, Y., Renuse, S., Chaerkady, R., Mathur, P. P., Ravikumar, R., and Pandey, A. (2012) Proteogenomic analysis of Candida glabrata using high resolution mass spectrometry. *J. Proteome Res.* **11,** 247–260
9. Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L., Muthusamy, B., Yadav, A. K., Shrivastava, P., Marimuthu, A., Anand, S., Sundaram, H., Kingsbury, R., Harsha, H. C., Nair, B., Prasad, T. S., Chauhan, D. S., Katoch, K., Katoch, V. M., Chaerkady, R., Ramachandran, S., Dash, D., and Pandey, A. (2011) Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. *Mol. Cell. Proteomics* **10,** M111.011627
10. Pawar, H., Sahasrabuddhe, N. A., Renuse, S., Keerthikumar, S., Sharma, J., Kumar, G. S., Venugopal, A., Sekhar, N. R., Kelkar, D. S., Nemade, H., Khobragade, S. N., Muthusamy, B., Kandasamy, K., Harsha, H. C., Chaerkady, R., Patole, M. S., and Pandey, A. (2012) A proteogenomic approach to map the proteome of an unsequenced pathogen - Leishmania donovani. *Proteomics* **12,** 832–844
11. Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. K., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. H., Gowda, H., and Pandey, A. (2014) A draft map of the human proteome. *Nature* **509,** 575–581
12. Gupta, N., Bandeira, N., Keich, U., and Pevzner, P. A. (2011) Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **22,** 1111–1120
13. Blakeley, P., Overton, I. M., and Hubbard, S. J. (2012) Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* **11,** 5221–5234
14. Trapnell, C., Pachter, L., and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111
15. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–578
16. Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., and Li, W. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41,** e74
17. Amanchy, R., Kalume, D. E., and Pandey, A. (2005) Stable isotope labeling with amino acids in cell culture (SILAC) for studying dynamics of protein abundance and posttranslational modifications. *Sci. STKE* **2005,** l2
18. Taouatas, N., Altelaar, A. F., Drugan, M. M., Helbig, A. O., Mohammed, S., and Heck, A. J. (2009) Strong cation exchange-based fractionation of Lys-N-generated peptides facilitates the targeted analysis of post-translational modifications. *Mol. Cell. Proteomics* **8,** 190–200
19. Kall, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **7,** 29–34
20. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–342

21. Jiao, X., Sherman, B. T., Huang da, W., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **28,** 1805–1806
22. Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., Maccoss, M., and Bafna, V. (2014) Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* **13,** 21–28
23. Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* **14,** 178–192
24. Letunic, I., Doerks, T., and Bork, P. (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40,** D302–D305
25. Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41,** D1063–D1069
26. Zheng, W., Xu, H., Lam, S. H., Luo, H., Karuturi, R. K., and Gong, Z. (2013) Transcriptomic analyses of sexual dimorphism of the zebrafish liver and the effect of sex hormones. *PLoS One* **8,** e53562
27. Ulitsky, I., Shkumatava, A., Jan, C. H., Subtelny, A. O., Koppstein, D., Bell, G. W., Sive, H., and Bartel, D. P. (2012) Extensive alternative polyadenylation during zebrafish development. *Genome Res.* **22,** 2054–2066
28. Nolte, H., Konzer, A., Ruhs, A., Jungblut, B., Braun, T., and Kruger, M. (2014) Global protein expression profiling of zebrafish organs based on in vivo incorporation of stable isotopes. *J. Proteome Res.* **13,** 2162–2174
29. Peri, S., and Pandey, A. (2001) A reassessment of the translation initiation codon in vertebrates. *Trends Genet.* **17,** 685–687
30. Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., Fan, L., Sandelin, A., Rinn, J. L., Regev, A., and Schier, A. F. (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* **22,** 577–591