TECHNIQUES

Piloting the Zebrafish Genome Browser

Anthony DiBiase,[1][*][†] Rachel A. Harte,[2][†] Yi Zhou,[1] Leonard Zon,[1] and W. James Kent[2]

This correspondence is a primer for the zebrafish research community on zebrafish tracks available in the UCSC Genome Browser at http://genome.ucsc.edu based on Sanger's Zv4 assembly. A primary capability of this facility is comparative informatics between humans (as well as many other model organisms) and zebrafish. The zebrafish genome sequencing project has played important roles in mutant mapping and cloning, and comparative genomic research projects. This easy-to-use genome browser aims to display and download useful genome sequence information for zebrafish mutant mapping and cloning projects. Its user-friendly interface expedites annotation of the zebrafish genome sequence. *Developmental Dynamics* 235:747–753, 2006. © 2005 Wiley-Liss, Inc.

INTRODUCTION

The Trans-NIH Zebrafish Genome Project Initiative has started to provide useful tools for researchers. International collaboration and foresight (especially the Wellcome Trust Sanger Institute and UCSC's Genome Informatics Group) have rapidly developed the capability for researchers to view the zebrafish genome.

The best zebrafish genome assemblies available today still exhibit a high degree of mis-assembly even in Zv5, which has just been released and is downloadable from http://www.sanger.ac.uk/Projects/D_rerio/Zv5_assembly_information.shtml). Zv4 (a major re-assembly when compared to Zv3) is available today as *danRer2*. It is this assembly that will be the focus of this report. Truth be told, the 1.56 gigabases of sequence (about $5.7\times$ coverage) in Zv4 assembly is still preliminary. Zv5 has 1.63 gigabases of sequence with a coverage of 6.5–$7\times$ as more reads have been sequenced (Sanger Institute,

---

ABBREVIATIONS danRer2 UCSC danio rerio (zebrafish) build 2 **BLAST** Basic local alignment sequence tool **BLAT** BLAST-like Alignment Tool **chr[1–25, Un, NA, M]** Zebrafish chromosome names: chr1-chr25, chrUn, chrNA, chrM **contig** A continuous sequence of DNA that has been assembled from overlapping cloned DNA fragments **FASTA** Text file format commonly used as input to BLAST **FPC** fingerprinted contig **GUI** graphical user interface **PSL** text file format for representing sequence alignments **SNPs** single nucleotide polymorphisms **STS markers** Sequence Tagged Site marker represents a single, unique, sequence-defined point in a genome **supercontig** A supercontig consists of one or more sequence contigs known to occur in a specific order and orientation **TIGR** The Institute for Genome Research (now part of the Craig Venter Institute) **Track mode** The UCSC genome browser displays tracks in one of the following modes: *Hide:* the track is not displayed at all. *Dense:* the track is displayed with all features collapsed into a single line. *Full:* the track is displayed with each annotation feature on a separate line. *Squish:* the track is displayed with each annotation feature shown separately, but at 50% the height of full mode. *Pack:* the track is displayed with each annotation feature shown separately and labeled, but not necessarily displayed on a separate line **Vega** Sanger's Vertebrate Genome Annotation database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence **WGS** whole genome shotgun sequencing method **WZ EST clusters** Washington University at St. Louis zebrafish EST clusters **ZGC** NCBI zebrafish genome collection **Zv5** Sanger's Zebrafish assembly version 5.

---

[1]Division of Hematology/Oncology, Children's Hospital, Karp Research Laboratories, Boston, Massachusetts
[2]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California
[†]Anthony DiBiase and Rachel A. Harte contributed equally to this work.
*Correspondence to: Division of Hematology/Oncology, Children's Hospital, Karp Research Laboratories, Boston, MA 02115. E-mail: adibiase@enders.tch.harvard.edu

2004). At the top level, the **danRer2** assembly is organized as:

chr1 to chr25: finished clones matched to WGS supercontigs.
chrNA: WGS contigs that could not be related to any FPC contig.
chrUn: WGS supercontigs that mapped to FPC contigs, unknown chromosome.
chrM: mitochondrion genome sequence was obtained from NCBI.

Having put the browser infrastructure and development resources in place, our team is ready to greet each new, improved release of the zebrafish assembly and quickly process it for the research community's use as part of the UCSC Genome Browser (Kent et al., 2002).

This primer is organized to allow researchers to rapidly perform informatics analyses. We first present the basic elements of the graphical user interface (GUI). Next, is a brief discussion of the available genomic tracks (viewable using the browser GUI and represented as colored, collinear blocks with text labels and strand annotation) and their salient characteristics. Next, we present some cookbook recipes for tasks that researchers routinely perform. We close with a glimpse into the next release of the browser.

BROWSER GUI AND TRACKS

The zebrafish browser may be accessed through the Gateway page for the genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway). Zebrafish should be selected from the pulldown list of genomes and an assembly date may be chosen from the assembly menu. Currently, the Zv4 (UCSC name: **danRer2**) Zebrafish assembly is available (June 2004), although the next freeze of the assembly (to be called **danRer3**) may be accessed in the near future. A position may be added to the position box and this may be a chromosome range, e.g., chr1 for the whole of chromosome 1 and chr1: 1,000–4,000 for the region from 1–4,000 on chromosome 1 (Fig. 1). After submitting this information, the default browser view for the chosen position will be displayed. Alterna-tively, this box may be used to search for accession names, gene names, or names of scientists who deposited sequences in GenBank. This is outlined on the Gateway page.

The default track display may be altered using the track display controls in the bottom half of the browser page and by clicking on the "configure tracks and display" button. Pull-down menus set the display mode for each track and are explained at the link http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#TRACK_CONT.

All tracks have hide, dense, and full modes while some additionally possess squish and pack modes. Clicking on a the track title above the track controls or the blue or gray bar perpendicular to a track on the track display takes the user to the page describing the track, its creation, data sources, and credits. Some tracks, such as the ESTs tracks, have filters on this page that allow the user to select features with certain characteristics, e.g., organism, author, tissue of origin, etc. Clicking on a feature within a track will bring up a details page with additional feature-specific information. This may include links to alignments or to additional external information depending on the type of track. Other tracks have the facility to obtain the DNA from the region of the feature with such options as repeat masking and for acquiring the DNA with additional upstream and downstream regions and the sequence may be reverse complemented. The DNA link on the blue bar at the top of the browser also performs a similar function. For the tracks belonging to the Genes and Gene Predictions Group, there is the ability to select 5′ UTR, coding region (CDS) or 3′ UTR exons for genomic DNA. The mRNA sequence and predicted protein sequence are also available. The zebrafish (**danRer2**) release includes the following tracks:

Mapping and Sequencing: Position, Contigs, Scaffolds, Radiation Hybrid Map, BAC Ends, Gap, GC Percent, Short Match, Restriction Enzymes.

Genes and Gene Prediction: RefSeq Genes, ZGC Genes, Ensembl Genes.

mRNA and EST: Zebrafish mR-NAs and ESTs, spliced ESTs, ZFish WZ EST Clusters (clustered ESTs from Washington University, St. Louis), Non-Zebrafish mRNAs, TIGR Gene Index.

Expression and Regulation: Affy Zebrafish Genechip: alignment of sequences used for probe design.

Comparative Genomics: Human, Mouse, Opossum, Fugu and Tetraodon Chain/Human, Mouse, Opossum, Fugu and Tetraodon Net/6- Way Conservation and Most Conserved (Zebrafish/Tetraodon/Fugu/Human/Mouse/Opossum multiple alignment and conservation)/ Human Proteins (tBLASTn of Human Known Genes).

Variation and Repeats: RepeatMasker/Simple Repeats.

Links to various tools reside on the blue bar along the top of the browser. BLAT is a very useful tool for aligning sequences to the genome. It is a tool similar to BLAST but allows rapid alignment against very large sequences such as genomes (Kent et al., 2002). The "Tables" link leads to the Table Browser where regions of the genome may be selected and data may be filtered and downloaded from the underlying database tables that contain the browser data (Karolchik et al., 2004). The "PDF/PS" tool produces a PDF or PostScript (PS) file of the image of the current view in the browser. The "Help" link provides information on getting started with using the Genome and the Table Browsers.

Downloads of sequence and alignment data for the whole genome may be obtained from the "Downloads" link on the blue bar at the left side or from: http://hgdownload.cse.ucsc.edu/downloads.html

Selecting the Zebrafish link takes the user to a list of available downloads. The Full data set includes the repeat masked (using RepeatMasker and Tandem Repeat Finder) genome sequence with repeats either in lowercase or replaced by capital N's. Different bioinformatics software may require different formats so we provide these two formats. The lowercase masked sequence is also available by individual chromosome from the "Annotation data by chromosome" link. The zebrafish mR-NAs, ESTs and RefSeqs, and non-zebrafish mRNAs are available for download from the full data set. The "Annotation database" link provides the data and a means of recreating the zebrafish database that is behind the ge-

nome browser. BLASTz (Schwartz et al., 2003) Fugu, human, and mouse alignments versus zebrafish are also available together with the corresponding chains and nets data.

The "Comparative Genomics" link at http://zfrhmaps.tch.harvard.edu/ZonRHmapper also displays reciprocal pre-BLAST results between human known genes and Fugu proteins, and between zebrafish genome sequences and Fugu proteins. In addition, a BLAST (Altschul et al., 1990) search engine is available to meet specific needs.

RECIPES

General notes: Tracks are bold, e.g., **Human net**. GUI buttons are double-quoted, e.g. "submit." Genome builds are bold italics, e.g., ***danRer***. "Sample position queries" on the gateway page http://genome.ucsc.edu/cgi-bin/hgGateway describes a wide range of descriptors you can search for sequence data. Clicking on the blue or gray bar at the side of the track or on the track name above the track control results in a description of the track being displayed.

2. Downloading Sequence: DNA and Amino Acids

Two recipes for downloading single sequences using genomic descriptors follow.

Using a chromosome name and genomic index.

From the gateway page, type your positional info (say, chr22:12879-24844) into the "position" box, then hit enter. Select the "DNA" link from the tracks page, which will take you to a page entitled "Get DNA in Window." Set the Sequence "Retrieval Region and Formatting Options," then click on the "Get DNA" button on the lower left.

Using a gene name.

From the gateway page, type your gene descriptor (say, NB) into the "position" box, then hit enter on your keyboard. You will then be in a page of hits based on your input. You must select (click on a link) for the hits you are specifically interested in. Choosing "BC059436" from the list of the hits under the heading "Zebrafish Aligned mRNA Search Results" takes

you to the tracks page. When you click on the colored descriptor text "gmb3" in the left-hand margin, you are taken to the page "RefSeq Gene." From here, you can download sequence for predicted protein, mRNA sequence, or genomic sequence.

Two recipes for downloading multiple sequences using genomic descriptors follow.

Download all BAC End Pairs in a genomic region using a list of constraints.

From the gateway page, click on the "Tables" link on the blue bar atop the page (Karolchik et al., 2004). Select "Mapping and Sequencing Tracks" from the "group" pull-down menu. Select "Radiation Hybrid Map" from the track pull-down menu. Select the "position" region radio button, and enter a position in the text box (say: chr19: 208146-227384). Type a name for your file in the "output file" text box, then click "Get Output."

Download all the exon peptide sequences for all zebrafish genes (known to date).

We will download an entire table from the database in this recipe. From the gateway page, click on the "Tables" link on the blue bar atop the page. Select "Genes and Gene predictions" from the "group" pull-down menu. Select either "RefSeq" or "Ensembl" from the track pull-down menu. Select the "genome" region radio button. Type a name for your file in the "output file" text box, then click "Get Output."

BLATing Sequence

Click on the "BLAT" link on the blue bar atop the Gateway (or any other) page.
Input sequence text: Cut FASTA sequence text of interest from a source and paste it into large input text box.
Input sequence from file: Type in or browse to the FASTA file name containing your sequence.

Click on "submit." A "BLAT Search Results" window will appear detailing alignments. Clicking on a "browser" link from this page takes you to the position in the genome with your current track display configuration. Clicking on a "details" link from this

page opens a window whose title is similar to "Alignment of danRer2_refGene_NM_213308 and chr1: 43721385-43722734."

BLAT is a BLAST-like Alignment Tool, which is fast and suitable for aligning sequences to a very large sequence such as a genome (Kent, 2002). The BLAT link may be found on the blue bar of the Index (http://genome.ucsc.edu), Gateway, and Zebrafish browser pages. First select the zebrafish genome from the "Genome" and "Assembly" menu pull-downs on the BLAT page. The query type may be selected as DNA, protein, translated RNA, or translated DNA. "BLAT's guess" is the default and it is good at distinguishing between a DNA and protein query. Generally, the hyperlink output is good if you wish to be able to view the alignment in the browser and also the aligned sequences. However, in order to obtain a text summary of the alignment and coordinates, a PSL output may be chosen. For more information see http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#BLATAlign.

Determining Human Homologs

Human homolog of zebrafish gene.

Configure the **Human Chain** track visibility to pack, then type the gene symbol of interest into the "position," hit enter on your keyboard. Click on the links from the details page. You will be directed to the corresponding regions of the human genome. A specific example for the twhh gene is shown in Figure 2a and b.

A user may have a zebrafish gene of interest for which they want to find the human homolog. It is possible to start either by searching for the zebrafish gene in the browser by searching using the gene name in the "Position" box or the user can BLAT a sequence of interest as described above. Once the gene is in the browser display, human homologs can be found by using the precomputed human comparison tracks. **Human Chain** and **Human Net** tracks are useful for viewing these comparative alignments and using the track controls, these may be switched to "full" visibility. The **Human Proteins**
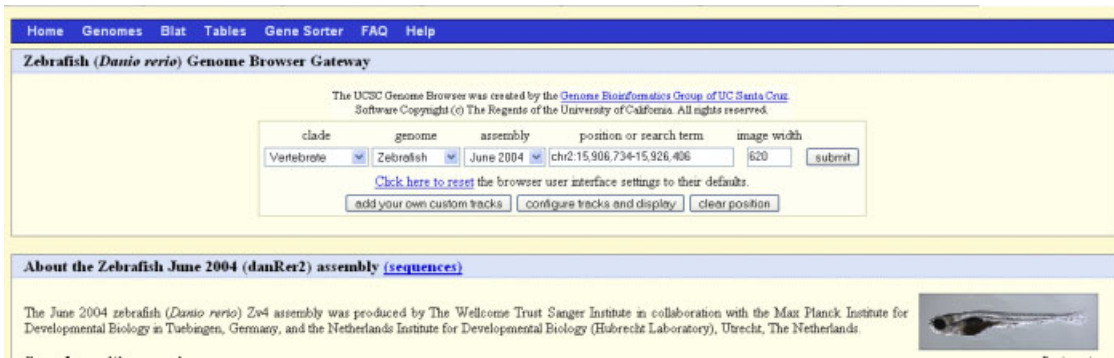
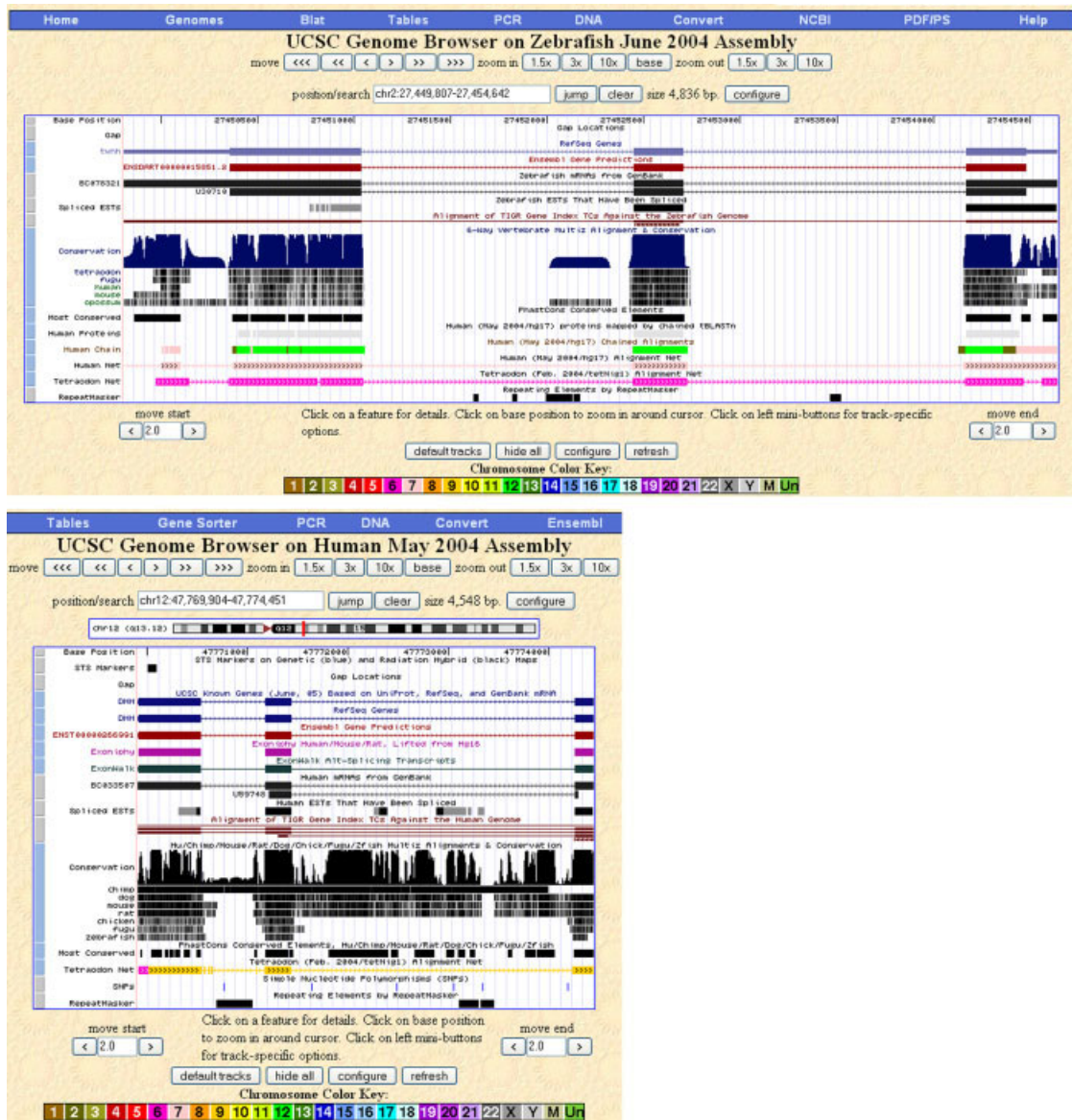**Fig. 1.** Zebrafish Genome Browser Gateway Page and its components.

Fig. 2.   a: Zebrafish Genome Browser in the region of the twhh gene. Coloring of alignments for the chain, net, and **Human Proteins** tracks represents the chromosome to which the region is aligned on the other organism's genome. The Chromosome Color Key is between the browser display and the track controls. Alignment chr12 + 47769k means this alignment is to the Human chromosome 12 on the + strand starting at around co-ordinate 47769k (exact co-ordinates are on the details page found by following the link from this label). chr 7 + 154749k aligns, in part, to the region of the human sonic hedgehog (shh) gene, chr 2 + 219745k aligns to the human indian hedgehog (ihh) gene, and chr7 + 47769k aligns to the human desert hegdehog (dhh) gene. **b:** Human Genome Browser in the region of the dhh gene. Alignment chr12 + 47769k from the zebrafish browser aligns to the region of the desert hedgehog homolog (dhh) gene on the human genome browser.

track started with predicted proteins from the **Known Genes mRNA** from human (**hg17** assembly, NCBI Build 34). Then, after identifying the exons by BLAT, the corresponding putative exons were found in the zebrafish genome using tBLASTn and, finally, these alignments were "chained" together to form longer alignments to determine gene structure. The human gene names are used as labels for these alignments in the zebrafish browser. On this page, there is a choice of coloring the alignments: (1) by score with shades of gray representing percent identity; (2) by chromosome color (the key for this is just below the browser track display) or (3) the alignments may be all displayed in black.

The **Human Chain** and **Human Net** tracks may be used in a similar way. BLASTz (Schwartz et al., 2003) is used to align two genomes and the chaining program chains together alignments to form gene structure. Clicking on these alignments will give information about the co-ordinates of the alignment on both genomes. There is also a link to the other organism's browser to show the region of alignment. With the **RefSeq** and **mRNA** tracks visible, it is possible to see if there are known homologs in this region. Alternatively, turning on some of the **Gene Prediction** or **EST** tracks would suggest whether there is evidence for genes in this particular region.

Determining Synteny

Gene synteny to human and mouse.

Configure **6-way Conservation, RefSeq Genes**, **Human Net**, and **Mouse Net** tracks to "full" mode (and any other species net tracks you are interested in). Type the gene symbol of interest into the "position" box, then type enter on your keyboard (if you don't have a gene name, you can enter a genomic position chrN: start – end). A new window displays a list of possible zebrafish gene candidates. Click on a desired link in this list. Use the zoom controls near the top of the page to focus. See the details below for content and explanations.

Syntenic regions can provide clues to how one genome evolved from another and so one can see the inversions, dele-
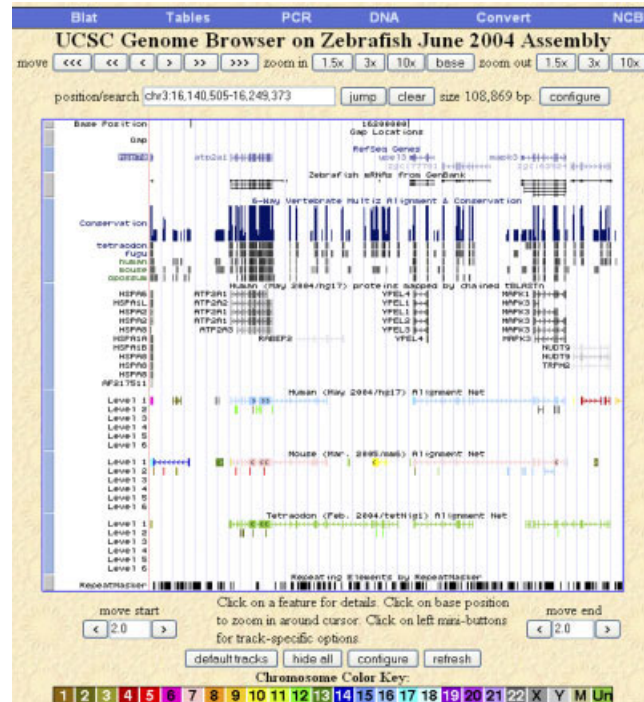

Fig. 3. Zebrafish Browser showing the Human Chains and the Human Net in the region of the zebrafish atp2a1, ypel3, mapk3, and zgc:77781 genes that have provisional status as RefSeqs.

tions, and which chromosomes or chromosomal regions are derived from each other. The chain and net tracks can be used in determining synteny between genomes (Kent et al., 2003). BLAT or a text search may again be used to initially identify the region of interest (see Fig. 3). BLASTz alignments are first chained to produce the gene structure and some low scoring chains are re-
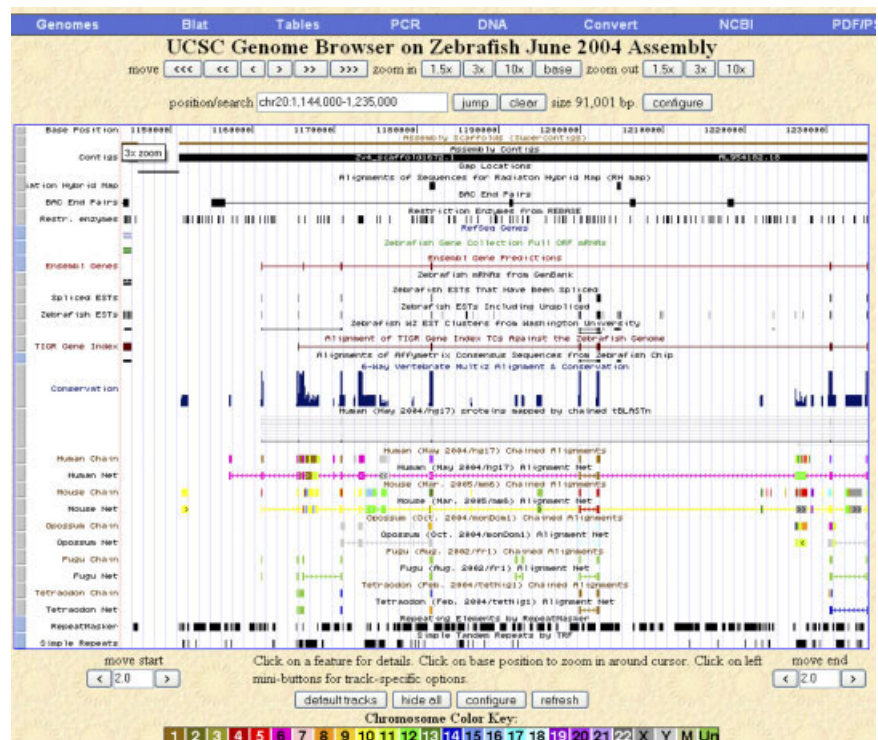

Fig. 4. Custom annotation tracks available in the *danRer2* build.

moved at this stage. The BLASTz blocks are in the same orientation and order in each species in order to be chained. In the production of the net track (see **Human Net** track in Fig. 2a), the highest scoring alignment is chosen and this is displayed at level 1 in this track. Only orthologous regions are shown.

If the position, chr3:16,170,001-16,270,001, is copied into the position box and the **Human**, **Mouse**, and **Tetraodon Net** tracks turned to "full" as in Figure 3, it can be seen that there are alignments in levels 1 and 2 of these net tracks. The top level (1) is the largest, highest scoring chain in this region. The boxes represent un-gapped alignment while the lines represent gaps and arrows show the direction of the alignment on the query genome (human in this case). Clicking on a line displays details about that gap, while clicking on a box gives details about the alignment with a link to show the actual alignment or to link to the corresponding region in the genome browser for the query. The **RefSeq** and **mRNA** tracks show the gene structure for the genes in zebrafish in this region: there are four genes with annotations (atp2a1, ypel3, mapk3, and zgc:77781). In level 1, the alignment is mainly light blue, showing that it is aligned to human chromosome 16. The equivalent regions for the mouse are pink indicating alignment to chr7 and for tetraodon the green shows alignment to chrUn_random, which are unmapped scaffolds. For **Human** and **Mouse Nets**, there is an alignment that corresponds to the position of the atp2a1 gene and another that contains the region of the ypel3, zgc:77781, and mapk3 genes. In all cases, these alignments are on the same chromosome for each species. Since these genes and alignments are found on the same chromosome within each species, they exhibit conserved synteny. For tetraodon, there is one alignment with the atp2a1 gene region, another with ypel3 and zgc:7778, and a third with the mapk3 gene region. These are all on chrUn_random in tetraodon and if a block in the aligning region is clicked, the details page is displayed and a link can be followed to the corresponding region in the tetraodon browser. This shows that these three alignments are on the same scaffold so it is likely that they are from the same chromosome and also show conserved synteny. Gaps in this top level chain are filled in by other chains at level 2, which may also have gaps filled in by chains in level 3 but there are no chains for level 2 gaps in this case. Annotation of the level 2–6 alignments indicates whether these alignments are syntenic (Syn), inverted (Inv), or non-syntenic (Non-Syn) in relation to the gap in the level above. In Figure 3, there are several green alignments in level 2, which are also aligned to human chromosome 12, and these are non-syntenic since they do not align to the same chromosome as the gap above in level 1.

Finding Markers in Gene Locus

Download all RH markers in a genomic region using a list of constraints.

From the gateway page, click on the "Tables" link on the blue bar atop the page. Select "Mapping and Sequencing Tracks" from the "group" pull-down menu. Select "Radiation Hybrid Map" from the track pull-down menu. Select the "position" region radio button, and enter a position in the text box (say: chr19:208146-227384). Type a name for your file in the "output file" text box, then click "Get Output."

Translating Coordinates Between Assemblies

When a new assembly is released, the coordinates for many annotation features may change. A user may wish to be able to find features of interest in a new assembly and to be able to translate the co-ordinates from an older assembly to the current one. From the Index, Gateway, or Browser pages, the "Help" link can be selected. The "Convert" link from the blue bar across the top of the screen explains how to convert data between assemblies.

CUSTOM ANNOTATION TRACKS

Figure 4 details the custom annotation tracks available in the **danRer2** build. You can add your own tracks using the info at http://genome.ucsc.edu/goldenPath/customTracks/custTracks.html.

FUTURE TRACKS

In the future, new tracks of interest to the zebrafish research community will be added to the browser. Here are some possibilities:

Mapping and Sequencing: Update (Zv5 assembly).

Genes and Gene Prediction: Vega, KnownGenes.

Variations and Repeats: STS markers, SNPs.

Vega genes (http://vega.sanger.ac.uk) is a set of manually curated annotations from The Wellcome Trust Sanger Institute, Cambridge, United Kingdom. Annotations are produced at the clone level by similarity searches against DNA and protein databases as well as using ab initio gene predictions. Genome comparisons between evolutionarily closely related species are also used to extend annotations. All of the data are useful in adding gene structures, polyA features, and gene descriptions to the genome. The Known Genes track will be created for the Zv5 assembly (**danRer3**) and this will consist of protein-coding genes based on the Ensembl gene set (Curwen et al., 2004). Links to other data sources such as in situ hybridization images (at ZFIN, http://zfin.org), protein structures, will be available through the details pages for the genes.

We welcome and encourage suggestions for new and interesting tracks from our users. There is a genome browser mailing list (genome@soe.ucsc.edu) to which you may subscribe. Here, users may make suggestions, participate in discussions, or ask questions about using various features of the genome browser. Subscription may be set up from this site: http://www.cse.ucsc.edu/mailman/listinfo/genome. In addition, new features and releases are announced through the genome-announce mailing list to which a subscription may be set up from this site: http://www.soe.ucsc.edu/mailman/listinfo/genome.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M. 2004. The Ensembl automatic gene annotation system. Genome Res 14:942–950.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. 2004. The UCSC Genome Browser Database. Nucleic Acids Res 32:D493–D496.

Kent WJ. 2002. BLAT: The BLAST-like alignment tool. Genome Res 12:656–664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12:996–1006.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D, Kent WJ. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100:11484–11489.

Sanger Institute. 2004. Zebrafish Sequencing Group data. ftp://ftp.sanger.ac.uk./pub/sequences/zebrafish [last accessed December 2005].

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. Human-mouse alignments with BLASTZ. 2003. Genome Res 13:103–107.