

Module 1: Where do I start? (Sanger)

i - The Zebrafish Genome Sequencing Page

Aims

- Introduce Sanger web pages
- Give examples of the services provided
- Show ways to navigate around the Sanger pages

Introduction

In spring 2001 the Wellcome Trust Sanger Institute started sequencing the genome of the zebrafish (*Danio rerio*). The strategy being used is clone mapping and sequencing from BAC and PAC libraries complemented with a whole genome shotgun (WGS) assembly. Some of these clones are selected for sequencing based on their location in the tiling path of the physical map. The released zebrafish assembly is based on the integration of the available finished clones with the WGS assembly contigs. The assembly is automatically annotated using the Ensembl pipeline and can be browsed on the Ensembl site. Assemblies are released once or twice a year depending on the available data. The current assembly is Zv5, which was released on May 27th 2005. The assembly will eventually consist solely of finished clones, with no sequence from the WGS assembly.

Sequences from finished clones come through the sequencing pipeline on a daily basis. They currently cover around 50% of the estimated 1.6 Gb size of the zebrafish genome. In a collaboration with ZFIN, finished clones are manually annotated. The finished clones with manual annotation can be browsed via the Vega database. These data are updated regularly to reflect the changes in the physical map and to make public the annotation. In modules 2.(i) and 2.(ii) the structure of the data in Ensembl and Vega is discussed in more detail, and exercises show how to get the best out of these resources.

The *Danio rerio* Sequencing Project Page

The main gateway to all the information regarding the zebrafish genome project is:

http://www.sanger.ac.uk/Projects/D_rerio

On the left-hand side of this page there is a quick-access toolbar with links to the services offered, and on the right-hand side recent news related to the project is shown. The page is divided into five parts:

- FAQs and contact information
- clone mapping and sequencing
- assembly releases
- other services
- contacts and links

The screenshot shows the website for 'The Danio rerio Sequencing Project' at the Wellcome Trust Sanger Institute. The page features a navigation menu on the left, a main content area with sections for 'Clone mapping and sequencing', 'Current Sequencing Status', 'Assembly releases', 'Other services', and 'Contacts and links', and a 'More News' sidebar on the right. Annotations in yellow boxes with arrows point to specific features:

- Quick-access toolbar**: Points to the left-hand navigation menu.
- FAQs and contact**: Points to the 'FAQs - frequently asked questions' section.
- news**: Points to the 'More News' sidebar.
- clone mapping, sequencing and manual annotation**: Points to the 'Clone mapping and sequencing' section.
- assembly releases**: Points to the 'Assembly releases' section.

The 'Current Sequencing Status' table is as follows:

06-Jul-2005	Unfinished	Finished	Total
<i>Danio rerio</i>	706,863,516	857,102,509	1,563,966,025

Contacting us

The email address for any enquiry regarding the project is:

zfish-help@sanger.ac.uk

We also have a link to an FAQs page where a wide range of questions regarding the project are already answered.

Clone mapping and sequencing

This page lists all the relevant links to the zebrafish clones, from their mapping to the sequence. There are links to the Vega database, the FPC database (see module 2.v) and to a Blast server for searching all the available sequences from the project.

Assembly releases

This page has information about the current and previous assemblies with links to FTP sites from which the sequences can be downloaded. There is

also a link to the trace repository. This is a database that features traces from several projects including all the zebrafish reads used in the whole genome shotgun assembly. These databases can be searched for alignments using SSAHA (see module 2.v).

Other services

This section has a link to an online RepeatMasker server. This is based on the same Rebase database of zebrafish repeats used in the Vega/Ensembl analysis.

The screenshot shows the Zebrafish RepeatMasker Server interface. At the top, it says "The Wellcome Trust Sanger Institute". The main heading is "Zebrafish RepeatMasker Server". Below this, it explains that the service screens DNA sequences in FASTA format against a library of repetitive elements identified for zebrafish. It lists the masking libraries, including all repeats submitted to rebase and specific zebrafish repeats. A "SEQUENCE DATA" section contains a text area for pasting a sequence, a "Choose File" button, and a text input for accession numbers. A "RESULTS" section has "Mask" and "Reset" buttons. Annotations include a yellow box "paste sequence or..." pointing to the text area and another yellow box "enter filename" pointing to the "Choose File" button.

The output of this service returns the original sequence but repeats are masked by strings of Ns.