

## **Module 2: Maps and Genome Sequence (NCBI)**

### **iv - How Do I Find a Zebrafish Gene in the Genome?**

#### **Aims**

- Introduce search tools
- Suggest methods for text or sequence searches
- Provide example text and sequence searches
- Show alternative ways to identify genomic placement of your gene or cDNA

#### **Introduction**

Locating the placement of a gene on the genome can be accomplished through sequence comparisons or by position information of the gene or other related mapped objects, such as BAC clones, SNPs, ESTs or STS markers.

To compare two sequences by BLAST, select a BLAST tool from the BLAST home page (<http://www.ncbi.nlm.nih.gov/BLAST/>) or by choosing a BLAST database from the zebrafish BLAST page (<http://www.ncbi.nlm.nih.gov/genome/seq/DrBlast.html>).

Additional related records can be viewed by following the provided Links to Assembly, WGS Project and UniSTS records from the GenBank record.

Another option to view a gene in its genomic context is to align a representative mRNA to a genomic sequence by NCBI's Splign tool which produces spliced global alignments (<http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi>).

#### **Exercises:**

- 1. BLAST:** identifying genomic placement via sequence comparison using MegaBLAST
- 2. Zebrafish BLAST:** querying against the WGS contigs or genomic (reference) sequence
- 3. Splign:** aligning an mRNA against a genomic contig to view spliced global alignments

## 1. BLAST

From the BLAST home page, choose one of the BLAST tools to submit a search.

Search the trace archives with MegaBLAST to identify the most similar WGS, EST or other *Danio rerio* trace sequence.

NCBI → BLAST Latest news: 13 June 2005 : BLAST 2.2.11 released

**About**

- Getting started
- News
- FAQs

**More info**

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

**Software**

- Downloads
- Developer info

**Other resources**


- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

<p><b>Nucleotide</b></p> <ul style="list-style-type: none"> <li>Quickly search for highly similar sequences (megablast)</li> <li>Quickly search for divergent sequences (discontiguous megablast)</li> <li>Nucleotide-nucleotide BLAST (blastn)</li> <li>Search for short, nearly exact matches</li> <li>Search trace archives with megablast or discontiguous megablast</li> </ul>	<p><b>Protein</b></p> <ul style="list-style-type: none"> <li>Protein-protein BLAST (blastp)</li> <li>Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)</li> <li>Search for short, nearly exact matches</li> <li>Search the conserved domain database (rpsblast)</li> <li>Protein homology by domain architecture (cdart)</li> </ul>
<p><b>Translated</b></p> <ul style="list-style-type: none"> <li>Translated query vs. protein database (blastx)</li> <li>Protein query vs. translated database (tblastn)</li> <li>Translated query vs. translated database (tblastx)</li> </ul>	<p><b>Genomes</b></p> <ul style="list-style-type: none"> <li>Human, mouse, rat, chimp <b>NEW</b>, cow, pig, dog, sheep, cat</li> <li>Chicken, puffer fish, zebrafish</li> <li>Environmental samples</li> <li>Malaria</li> <li>Insects, nematodes, plants, fungi, microbial genomes, other eukaryotic genomes</li> </ul>
<p><b>Special</b></p> <ul style="list-style-type: none"> <li>Search for gene expression data (GEO BLAST)</li> <li>Align two sequences (bl2seq)</li> <li>Screen for vector contamination (VecScreen)</li> <li>Immunoglobulin BLAST (IgBlast)</li> <li>SNP BLAST</li> </ul>	<p><b>Meta</b></p> <ul style="list-style-type: none"> <li>Retrieve results</li> </ul>

**Organism-specific BLAST**

MegaBlast

 **megablast BLAST**

Nucleotide Protein Translations Retrieve results for an RID

### Trace Archive database Mega BLAST search

**Choose a Query sequence**

[Search](#)

Load query file from disk

[Set subsequence](#) From:  To:

[Choose database](#)

[Return alignment endpoints only](#)

**Submit for BLAST**

Now:  or

---

### Options

for advanced blasting

[Hits computed](#)

[Choose filter](#)  Low complexity  Rodent Repeats  Mask for lookup table only  Mask lower case

BLAST Results:



## results of BLAST

### BLASTN 2.2.10 [Oct-19-2004]

RID: 1119441055-19332-153186667233.BLASTQ2

**Database:** Danio\_rerio\_WGS  
22,634,736 sequences; 21,084,232,232 total letters

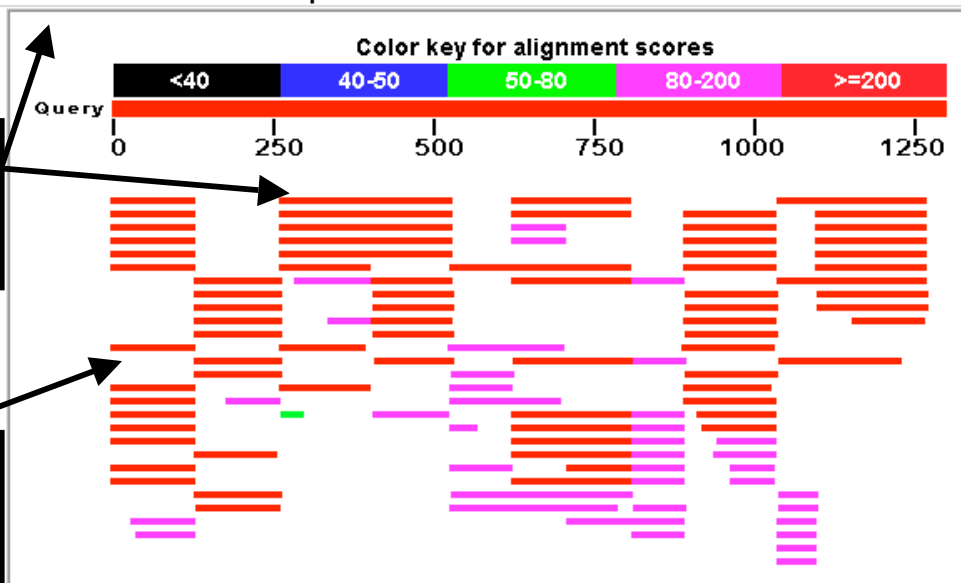
If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)

[Taxonomy reports](#)

**Query=** gi|67078407|ref|NM\_001024735.1| Danio rerio zgc:63996 (zgc:63996), mRNA  
(1293 letters)

### Distribution of 139 Blast Hits on the Query Sequence

611790792 zDH86-629e15.p1k S=234 E=7.1e-58



Mouse-over  
display to view  
Trace id

Click on any bar  
to jump to the  
alignment

```
>gnl|ti|611790792 zDH86-629e15.plk
      Length=846

Score = 272 bits (137), Expect = 3e-69
Identities = 140/141 (99%), Gaps = 0/141 (0%)
Strand=Plus/Minus

Query 264 AGATGTGGAGAGTGTGATGAACAGCATCGTGTCTCTGCTGCTGATTCTGGAGACGGAGAA 323
      |||
Sbjct 457 AGATGTGGAGAGTGTGATGAACAGCATCGTGTCTCTGCTGCTGATTCTGGAGACGGAGAA 398

Query 324 GCAGGAGGCTCTTATTGAAAGCTTATGTGAGAAGCTGGTGAAGTTTCGTGAGGGTGAACG 383
      |||
Sbjct 397 GCAGGAGGCTCTGATTGAAAGCTTATGTGAGAAGCTGGTGAAGTTTCGTGAGGGTGAACG 338

Query 384 GCCCTCGCTTCGGATGCAGCT 404
      |||
Sbjct 337 GCCCTCGCTTCGGATGCAGCT 317
```

View the alignment or follow the link to the trace archive

Click on this link to go to the trace archive record

Trace Archive for ti:611790792

(<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=retrieve&dopt=fasta&val=611790792>)

The Trace Archive consists of the raw, single-pass reads of DNA sequence generated from large-scale sequencing projects.

Home Search Site Map Main Statistics Tracking System Obtaining Data BLAST Documents

Searching Tips • Searchable Fields • Registered Species • Submitting Centers • FTP

Style: Blue Sky

Enter a query string or TI number

Search 611790792 alt

Save result of search as trace .tar  .gz file.

Save  All  FASTA  Quality  SCF  Info-XML  Info-Table  Mate Pair

Retrieve

Show as FASTA  in color  NULL values

Search result:

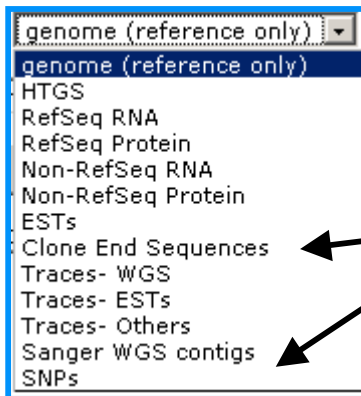
Your request is: 611790792

```
>gnl|ti|611790792 name:zDH86-629e15.plk mate:611686865 mate_name:zDH86-629e15.plk template:zDH86-629e15
CAGCTCGGTACCCCTTCATTTATTATCCATGATTAACCACAAAGACTATATACCTTCATATTTCAAACAC
AGTTTAATAAAATGACGGCGGTTTCTATACCTGGTCTAGGTCTGTAGGCATGAAGGTGATGGCGTTGCAG
GTTGCCCGCCACTTTAATCAGACTGCAGTACACAGTGTGTCTCACTGGAGTGTTCGTCGCATACCATGGA
ACAGATTACTCAAACCTAAGAGAAGCAATACGAGTTCAATACAAGCAAACATACAAATAATTAATACAAGGA
AATTAATAAAAAAAAAAAAAAAGTCTAGACACTCACAGCTGCATCCGAAGCGAGGGCCGTTTCCCTCAC
GAAACTTCAGCACTTCTACATAAGCTTTCAATCAGAGCCTCCTGCTTCTCCGTCTCCAGAATCAGCAG
CAGAGACACGATGCTGTTATCACAACCTCTCCACATCTGCAGACATCAGCAGAGGGGGACAAAAAAGGTT
ATTTGTGTATCCGATCTATAAAGAGAAGCTCTGATTAGATTCCGCCAGATATCAGTATTGGTCAATAATAA
AAAAATATTCTCTGACTAACAGATGTAGGACAAAAGGGGACACCAATAAGTTCAAGTAAAGGGGCCAAAAG
AGGGCTTCAGAAATTGCAATTAGTAAATGATTTTAGGAAAAGAAAGGTTCTAGATTTTTCTGTTCTGCAGTT
AAGGTTAACACATAACAATTGATCCGCTGTATTTGTGTTTTGTCTAATCCTTTTACATTTGTTCAITTCAG
CAGAAATCGTTTATACTATAAAAACTAAAAACAGAGATCACAAGGCAGCATTATATTGACCTTGTCTAAA
ATATGA
```

## 2. Zebrafish BLAST

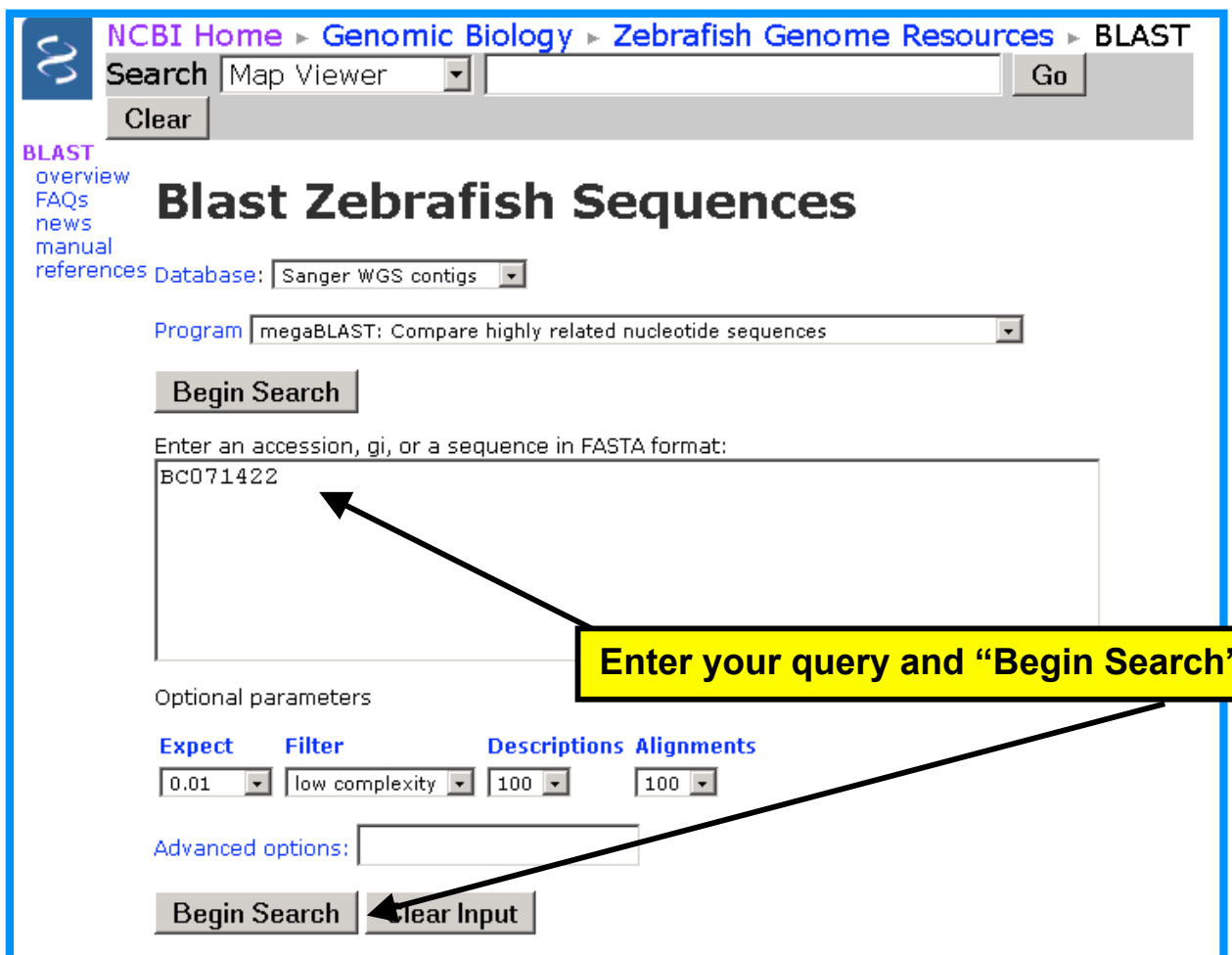
The Zebrafish BLAST page allows you to submit a query sequence against zebrafish-specific databases.

(<http://www.ncbi.nlm.nih.gov/genome/seq/DrBlast.html>)



genome (reference only) ▾  
genome (reference only)  
HTGS  
RefSeq RNA  
RefSeq Protein  
Non-RefSeq RNA  
Non-RefSeq Protein  
ESTs  
Clone End Sequences  
Traces- WGS  
Traces- ESTs  
Traces- Others  
Sanger WGS contigs  
SNPs

**Choose a sequence database:  
BLAST your ZGC clone against the  
Zebrafish Clone End Sequences or  
the Sanger WGS contigs**



NCBI Home ▸ Genomic Biology ▸ Zebrafish Genome Resources ▸ BLAST

Search  Map Viewer ▾

BLAST  
overview  
FAQs  
news  
manual  
references

## Blast Zebrafish Sequences

Database:

Program

Enter an accession, gi, or a sequence in FASTA format:

Optional parameters

Expect  Filter  Descriptions  Alignments

Advanced options:

**Enter your query and “Begin Search”**

BLAST Results



## results of BLAST

### BLASTN 2.2.10 [Oct-19-2004]

RID: 1119444995-16813-41293302633.BLASTQ4

**Database:** sanger\_wgs\_contigs  
21,333 sequences; 1,560,480,686 total letters

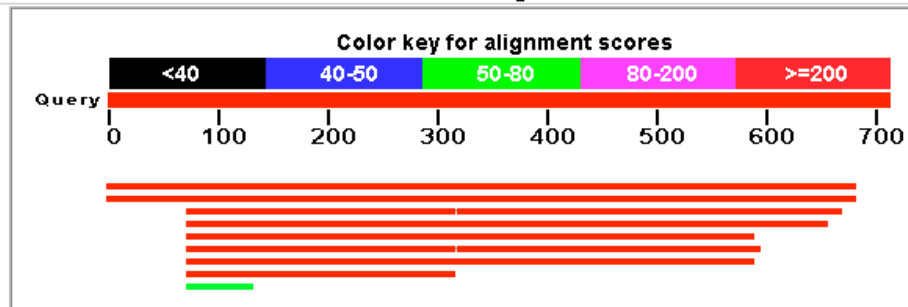
If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)

[Taxonomy reports](#)

**Query=** gi|47937949|gb|BC071422.1| Danio rerio zgc:86750, mRNA (cDNA clone MGC:86750 IMAGE:6899054), complete cds  
(714 letters)

### Distribution of 34 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



Sequences producing significant alignments:			Score (Bits)	E Value
<a href="#">emb CAAK01000032.1 </a>	Danio rerio whole genome shotgun, scaffol...		<a href="#">700</a>	0.0
<a href="#">emb CAAK01019643.1 </a>	Danio rerio whole genome shotgun, scaffol...		<a href="#">700</a>	0.0
<a href="#">emb CAAK01000072.1 </a>	Danio rerio whole genome shotgun, scaffol...		<a href="#">517</a>	1e-144
<a href="#">emb CAAK01000709.1 </a>	Danio rerio whole genome shotgun, scaffol...		<a href="#">469</a>	1e-130

Click to view GenBank record

Click on Score to view the alignment

Alignments:

## Alignments

```
>emb|CAAK01000032.1| D Danio rerio whole genome
genome shotgun sequence
Length=1718556
```

```
Score = 700 bits (364), Expect = 0.0
Identities = 366/367 (99%), Gaps = 0/367 (0%)
Strand=Plus/Plus
```

```
Query 318      GTACAGGGGATCCTACAGAATGAGGATCTACGAGAGGGACAACCTTCATGGGTCAGATGTA 377
                |||
Sbjct 205834    GTACAGAGGATCCTACAGAATGAGGATCTACGAGAGGGACAACCTTCATGGGTCAGATGTA 205893

Query 378      CGAGATGATGGATGACTGTGACAACATCATGAACCGTTACCGCATGTCTCACTGCCAGTC 437
                |||
Sbjct 205894    CGAGATGATGGATGACTGTGACAACATCATGAACCGTTACCGCATGTCTCACTGCCAGTC 205953

Query 438      CTGTCATGTGATGGATGGCCACTGGCTCTTTTATGACCAGCCCAACTACAGAGGCAGGAT 497
                |||
Sbjct 205954    CTGTCATGTGATGGATGGCCACTGGCTCTTTTATGACCAGCCCAACTACAGAGGCAGGAT 206013

Query 498      GTGGCACTTCGGGCCTGGGCAGTACAGGAACTTCAGCAATTATGGTGGCATGAGATTCAT 557
                |||
Sbjct 206014    GTGGCACTTCGGGCCTGGGCAGTACAGGAACTTCAGCAATTATGGTGGCATGAGATTCAT 206073

Query 558      GAGCATGAGGCGCATCATGGACTCTTGGTACTAGAATTTATTTGAATAAAAATACTTCTC 617
                |||
Sbjct 206074    GAGCATGAGGCGCATCATGGACTCTTGGTACTAGAATTTATTTGAATAAAAATACTTCTC 206133

Query 618      TAAGATATTAACATTGTCTTGAATATAATTAATGCCACTAACAATAAAAACAATATCCA 677
                |||
Sbjct 206134    TAAGATATTAACATTGTCTTGAATATAATTAATGCCACTAACAATAAAAACAATATCCA 206193

Query 678      CAAATAC      684
                |||
Sbjct 206194    CAAATAC      206200
```

Click to view the GenBank record and view related Assembly, WGS Project and

Click on the Links pull-down menu to follow the links to the related Assembly, WGS Project and UniSTS records

The screenshot shows the NCBI Entrez Nucleotide search interface. At the top, there are navigation tabs for PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, OMIM, and Books. The search bar contains "Nucleotide" and "for" with a "Go" button. Below the search bar are options for "Limits", "Preview/Index", "History", "Clipboard", and "Details". A "Display" button is set to "GenBank", and a "Send" button is set to "all to file". The search range is from "begin" to "end". There are checkboxes for "Reverse complemented strand" and "Features" (SNP, CDD, MGC, HPRD, STS). The search results show one entry: "1: CAAK01000032. Reports Danio rerio whole...[gi:58293633]". A "Links" menu is open, showing options for Assembly, WGS Project, Taxonomy, and UniSTS. The main content area displays the following information:

LOCUS CAAK01000032 1718556 bp DNA linear VRT 19-JAN-2005  
 DEFINITION Danio rerio whole genome shotgun, scaffold Zv4\_scaffold32, whole genome shotgun sequence.  
 ACCESSION CAAK01000032 [CAAK01000000](#)  
 VERSION CAAK01000032.1 GI:58293633  
 KEYWORDS WGS.  
 SOURCE Danio rerio (zebrafish)  
 ORGANISM [Danio rerio](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Danio.  
 REFERENCE 1  
 AUTHORS Jekosch, K., Caccamo, M., Ning, Z., Humphray, S., Scott, C., Barlow, K., Bradley, A., Burton, J., Clark, R., Elliot, D., Grafham, D., Hunt, A., Jones, M., Lloyd, D., Lloyd, C., Matthews, L., McLaren, S., McLay, K., Oliver, K., Palmer, S., Plumb, R., Quail, M., Riddle, C., Shownkeen, R., Sims, S., Threadgold, G., Willey, D., Windsor, C., Hubbard, T., Beck, S. and Rogers, J.

## Assembly: WGS Assembly Chromosome 1 Contig

NCBI Nucleotide

Search Nucleotide for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display GenBank Send all to file

Range: from begin to end  Reverse complemented strand Features:

SNP  CDD  MGC  HPRD  STS

1: [NW\\_633982](#). Reports ...[gi:67044306]

[Click here to see all features and the sequence of this contig record.](#)

LOCUS NW\_633982 1718556 bp DNA linear CON 08-JUN-2005

DEFINITION *Danio rerio* chromosome 1 genomic contig.

ACCESSION NW\_633982

VERSION NW\_633982.1 GI:67044306

KEYWORDS .

SOURCE *Danio rerio* (zebrafish)

ORGANISM [Danio rerio](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
Actinopterygii; Neopterygii; Teleostei; Ostariophysi;  
Cypriniformes; Cyprinidae; *Danio*.

COMMENT GENOME ANNOTATION [REFSEQ](#): NCBI contigs are derived from assembled genomic sequence data.  
Also see:  
[Documentation](#) of NCBI's Annotation Process

The DNA sequence is from the genome assembly released by the Wellcome Trust Sanger Institute as Zv4, 12 July 2004 (see [http://www.sanger.ac.uk/Projects/D\\_rerio/Zv4\\_assembly\\_information.shtml](http://www.sanger.ac.uk/Projects/D_rerio/Zv4_assembly_information.shtml)).

FEATURES Location/Qualifiers  
source 1..1718556  
/organism="Danio rerio"  
/mol\_type="genomic DNA"  
/strain="Tuebingen"  
/db\_xref="taxon:7955"  
/chromosome="1"

CONTIG join([BX119979.9](#):1..194023,[CAAK01000032.1](#):194024..649943,  
complement([BX511095.5](#):39775..154840),[CAAK01000032.1](#):765010..828010,  
[BX537295.6](#):1..183025,[CAAK01000032.1](#):1011036..1152097,  
[BX247884.6](#):1..108530,[CAAK01000032.1](#):1260628..1495069,  
complement([BX323083.6](#):56674..106933),  
[CAAK01000032.1](#):1545330..1546329,[BX119983.4](#):1..172227)

//

View NCBI annotation documentation

View the assembly instructions for the Chromosome 1 contig and follow links to clone and WGS sequences

UniSTS: view markers placed on the genomic contig by ePCR

NCBI **UniSTS** Integrating Markers and Maps My NCBI [Sign In] [Register]

PubMed All Databases BLAST OMIM Taxonomy Structure

Search UniSTS for  Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Send to

All: 88 in\_Gene: 43 Mapped: 5 OMIM: 0 with\_SNP: 2

Items 1 - 20 of 88 Page 1 of 5 Next

- 1: [UniSTS:190496](#) Links  
**fb36h10.x1**  
*Danio rerio* locus  
Found by e-PCR in sequences from *Canis familiaris*, *Danio rerio*, *Mus musculus* and *Rattus norvegicus*.
- 2: [UniSTS:202319](#) Links  
**fa11f10.s1**  
*Danio rerio* locus wdhd1  
Found by e-PCR in sequences from *Danio rerio*.
- 3: [UniSTS:193148](#) Links  
**fc23d08.x1**  
*Danio rerio* multiple loci  
Found by e-PCR in sequences from *Danio rerio*, *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*.

About Entrez

Entrez UniSTS

Help

Query tips

Submit

Submit map

FTP site

Statistics

Related sites

e-PCR

Map Viewer

Gene

UniGene

dbSNP

GeneMap'99

GDB

MGI

RGD

RHdb

**Click on the UniSTS id to view the e-PCR results and links to Gene and MapViewer**

**Go to UniSTS for fa11f10.s1**

UniSTS: (<http://www.ncbi.nlm.nih.gov/genome/sts/sts.cgi?uid=202319>)

<p>Entrez UniSTS</p> <p>Help</p> <p>Query tips</p> <p>Submit</p> <p>Submit map</p> <p>FTP site</p> <p>Statistics</p> <p>Related sites</p> <p>e-PCR</p> <p>Map Viewer</p> <p>Gene</p> <p>UniGene</p> <p>dbSNP</p> <p>GeneMap'99</p> <p>RHdb</p> <p>GDB</p> <p>MGD</p> <p>ZFIN</p> <p>Genomic biology</p> <p><i>Bos taurus</i></p> <p><i>Canis familiaris</i></p> <p><i>Danio rerio</i></p> <p><i>Homo sapiens</i></p> <p><i>Mus musculus</i></p> <p><i>Rattus norvegicus</i></p> <p><i>Sus scrofa</i></p>	<p>UniSTS:202319 <span style="float: right;"><a href="#">Links</a></span></p> <p><b>fa11f10.s1</b></p>
	<p><b>Primer Information</b> <span style="float: right;">?</span></p>
	<p>Forward primer: <b>GACTACTGGGTTCAGAAATGGG</b></p> <p>Reverse primer: <b>TTGAAAGCAGTTCCTGTCTCGC</b></p> <p>PCR product size: <b>121 (bp), <i>Danio rerio</i></b></p>
	<p><b><i>Danio rerio</i></b></p>
	<p><b>Name:</b> fa11f10.s1</p> <p><b>Also known as:</b> FA11F10.S1</p>
	<p><b>Cross References</b> <span style="float: right;">?</span></p>
	<p><b>Gene</b>    GeneID: <a href="#">322945</a></p> <p>             Symbol: <a href="#">wdhd1</a></p> <p>             Description: WD repeat and HMG-box DNA binding protein 1</p> <p>             Position:</p> <p><b>UniGene</b> <a href="#">Dr.33169</a>    WD repeat and HMG-box DNA binding protein 1</p> <p>   <a href="#">Dr.425</a>            Zgc:92585</p>
	<p><b>Electronic PCR results</b> <span style="float: right;">?</span></p>
	<p><b>ESTs (4)</b></p> <p><a href="#">AA494843.1</a>                            114 .. 234</p> <p><a href="#">BI980695.1</a>                            119 .. 239</p> <p><a href="#">BM184900.1</a>                            247 .. 367</p> <p><a href="#">BQ092342.1</a>                            110 .. 230</p>
	<p><b>Whole Genome Shotgun sequences (1)</b></p> <p><a href="#">CAAK01000032.1</a>                    508682 .. 508802</p>

### 3. Splign

View the spliced alignment of your ZGC clone against the Sanger WGS contig.

The screenshot shows the Spleign web interface. At the top left is the NCBI logo. To its right is the Spleign logo, which includes a colorful bar chart. Below the logos is a navigation menu with links for HOME, SEARCH, SITE MAP, Overview, Online, Download, Documentation, and Contacts. The main content area contains instructions: "Please specify input sequences by GI/Accession or in FASTA format. Examples (click to select):" followed by a list of four examples. Below this are two input fields: "cDNA:" with the value "NM\_001024735" and "Genomic:" with the value "NM\_639445". To the right of the Genomic field are "From:" and "To:" input boxes with values "1" and "max" respectively. Below the input fields are two checkboxes: "Reverse and complement the query (cDNA)" and "Cross-species mode". To the right of these checkboxes is a "Browse..." button. At the bottom left is an "Align" button. Three yellow callout boxes with black arrows point to specific parts of the interface: one points to the "Select query and target sequences" text, another points to the "Cross-species mode" checkbox, and a third points to the "Align" button.

NCBI

Spleign

HOME SEARCH SITE MAP Overview Online Download Documentation Contacts

Please specify input sequences by GI/Accession or in FASTA format.  
Examples (click to select):

- AB010263 / NT\_033778 (1 model, generic)
- AF034611 / NT\_077569 (1 model, 67 exons)
- NM\_020978 / NG\_004750 (6 models, unaligned segment)
- AF238306 / NT\_033777 (1 model, many frameshifts)

cDNA: NM\_001024735

Genomic: NM\_639445

From: 1 To: max

Reverse and complement the query (cDNA)

Cross-species mode

Browse...

Align

Select query and target sequences

You can also reverse and complement the query sequence or run the comparison in cross-species mode

**Alignment Results:**  
Click on each exon to view alignment statistics, cds alignment (pink) and splice sites (green)

