NCBI Molecular Biology Resources

PRIMO12: Short Course in Aquatic Toxicology

Peter Cooper
National Center for Biotechnology Information

May 9, 2003

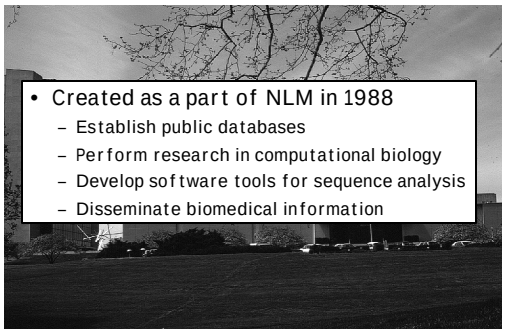
NCBI

Talk Outline

- About NCBI
- NCBI Databases
  - zebrafish sequences
- Integrated Access
  - Entrez
  - BLAST

NCBI

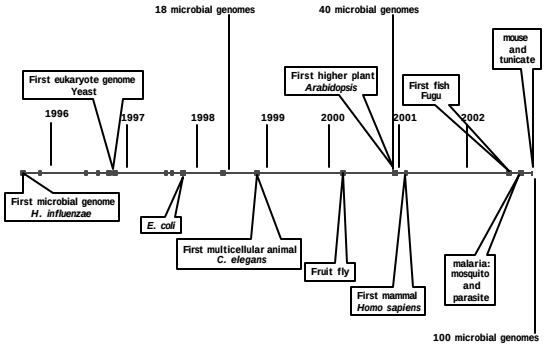
The National Center for Biotechnology Information



- Created as a part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information

NCBI

The Genome Sequencing Era



NCBI

Complex Genomes Jan. 2003

- Chordates
  - Human
  - Mouse
  - Rat
  - Pufferfish
  - Sea squirt (Ciona)
- Arthropods
  - *D. melanogaster*
  - *D. simulans*
  - *A. gambiae*
- Higher plants
  - *Arabidopsis*
  - Rice
- Fungi
  - *Aspergillus terreus*

NCBI

Coming soon …

- In progress
  - purple sea urchin
  - zebrafish
- NHGRI's Priority Organisms
  - Chicken
  - Cow
  - Dog
  - Chimpanzee
  - Honeybee
  - *Tetrahymena*
  - *Oxytichia*
  - Several fungi
- Over 100 bacterial genomes in progress

NCBI

```
103621 aataaaatac acactcacgc acacacatat gtactatata tatatatata tatatatata
103681 tatatatata tatatatata tatatttaaa gagtgtgtgt gtgtgtgtgt gtgtgtgtgt
103741 gtgtgtgagt gtgtgtgtgt gtgtgtgtgt gtgtgtgtgt gtctagatag agagagagag
103801 agagagagag agaataaata aaaaaaaaaa agaaagagag agagagagtg tgtgcgtgtg
103861 tgtgtgtgtg tgtgtgtgtg tatatagaga tatacatata gacagagaga gagagtgtgt
103921 gtatatatat atatatatat atatatatat agannnnnnn nnnnnnnnn nnnnnnnnn
103981 nnnnnnnnnn nnnnnnnnn nnnnnnnnn nnnnnnnnn nnnnnnnnn nnnnnnnnn
104041 nnnnnnnnn nnnagtaaca gtacaagtat gatgtgatgt agttagtaaa gtaagcagtc
104101 agacgacatc atattgtttg tttattctaa atatcttttt gagggggcac ttgattaagc
104161 agaatggaaa gaaacatggc ttacgattta tatagactag gataacgttt ttctctcacg
104221 cttgagccac agatctgaca gattatagca gctttaacac acatttcaaa gttgtgtgtc
104281 ataaaaacac tctgtaattc actcaaaatg ctattgatac ccattttcgg atcacaaatt
104341 accggt                                                gctggtcact
104401 atggct                                                tgggggtgtg
104461 aaccta                                                atcggtttaa
104521 ttcgatatct cggtgcatct cagtgcattg aagatgcttt ccatacataa ttagattttt
104581 tcttcacaac ataacaattt ttttattcaa atctatttat atatatatat ccagggatgc
104641 catctgggcc cgtagctttt cggatgttta cgcgacaatc aatagatcta gttactttca
104701 ctacagaaac agagagtgtg ctggcattta caatatcggc cgtgggagcg tgcagtgtgt
104761 cacgctgtcc tgggcttcaa agcaggcatg gaacgagttg agctcgtccg gcagagatgc
104821 aacaatgttt acagtggcgg gtttcttact tttaaagtct gtgatttgat taggtctttg
104881 ccacatgctc cttgcgttat tggtgttaaa ttgaccctct accttgtctc tgtattgacg
104941 ttttgatgct ttgacagttt ttccaaggct gtaatttgct cgtttttgtt cctcagcatt
105001 cctggattta taggtggaag ctcgcactga tcgggttgtt caaacttcac tattgatcca
105061 tggtttttgg ttggggtaga cacggacagt ttttgttggc accacgtcgt ctatacattt
105121 tctgatgaag cacgtaactg tgtctgagta cacttcgatg tcatcatcag aagctacctg
```

NCBI Sequence Databases

NCBI

Zebrafish Nucleotide Sequences

Display Summary Show 20 Send to: text

Items 1-20 of 482332

1: CB891047
STH00833 oligo dT primed shield stage Danio rerio cDNA
gi|30098216|gb|CB891047.1|[30098216]

2: CB891046
STH00832 oligo dT primed shield stage Danio rerio cDNA
MRNA sequence
gi|30098215|gb|CB891046.1|[30098215]

3: AC142089
Danio rerio clone CH211-4J20 strain Tue, WORKING DR
gi|30089719|gb|AC142089.2|[30089719]

4: AC139110
Danio rerio clone CH211-154N18 strain Tue, WORKING D
pieces
gi|30089718|gb|AC139110.3|[30089718]

5: AC139622
Danio rerio clone CH211-11P9 strain Tue, WORKING DR
gi|30089717|gb|AC139622.3|[30089717]

6: NM_178223
Danio rerio mitogen-activated protein kinase 14b (mapk14-b
gi|30089691|ref|NM_178223.1|[30089691]

7: NM_178222
Danio rerio single-minded homolog 1 (Drosophila) (sim1), r
gi|30089689|ref|NM_178222.1|[30089689]

8: BX000363
Danio rerio clone CH211-10K11, *** SEQUENCING IN P

29662: BQ783369
fab29e04.y1 Zebrafish SJD 15 day embryo Danio rerio cDNA clone IMAGE:5904654 5', MRNA sequence
gi|21991841|gb|BQ783369.1|[21991841]

29663: BQ783368
fab29a03.y1 Zebrafish SJD 15 day embryo Danio rerio cDNA clone IMAGE:5904652 5' similar to TR:Q9QXP0 Q9QXP0 RH TYPE C GLYCOPROTEIN. .; MRNA sequence
gi|21991840|gb|BQ783368.1|[21991840]

29664: BQ783367
fab29a01.y1 Zebrafish SJD 15 day embryo Danio rerio cDNA clone IMAGE:5904648 5' similar to SW:COXG_BOVIN P00429 CYTOCHROME C OXIDASE POLYPEPTIDE VIB .; MRNA sequence
gi|21991839|gb|BQ783367.1|[21991839]

29665: AF434664
Danio rerio insulin-like growth factor binding protein-1 mRNA, complete cds
gi|21954646|gb|AF434664.1|[21954646]

Traditional GenBank Records

29666: AF330001
Danio rerio UNC45-related protein mRNA, complete cds
gi|18033184|gb|AF330001.1|[18033184]

29667: AF469651
Danio rerio Na,K-ATPase beta subunit isoform 3a (atp1b3a) mRNA, complete cds
gi|18766820|gb|AF469651.1|[18766820]

29668: AF527755
Danio rerio Cyp11a1 mRNA, complete cds
gi|21913314|gb|AF527755.1|[21913314]

NCBI

Zebrafish Sequences: Bulk

Display Summary Show 20 Send to: text

Items 1-20 of 482332

1: CB891047
STH00833 oligo dT primed shield stage Danio rerio cDNA
gi|30098216|gb|CB891047.1|[30098216]

EST

2: CB891046
STH00832 oligo dT primed shield stage Danio rerio cDNA
MRNA sequence
gi|30098215|gb|CB891046.1|[30098215]

EST

3: AC142089
Danio rerio clone CH211-4J20 strain Tue, WORKING DR
gi|30089719|gb|AC142089.2|[30089719]

4: AC139110
Danio rerio clone CH211-154N18 strain Tue, WORKING D
pieces
gi|30089718|gb|AC139110.3|[30089718]

HTG

5: AC139622
Danio rerio clone CH211-11P9 strain Tue, WORKING DR
gi|30089717|gb|AC139622.3|[30089717]

HTG

6: NM_178223
Danio rerio mitogen-activated protein kinase 14b (mapk14-b
gi|30089691|ref|NM_178223.1|[30089691]

7: NM_178222
Danio rerio single-minded homolog 1 (Drosophila) (sim1), r
gi|30089689|ref|NM_178222.1|[30089689]

8: BX000363
Danio rerio clone CH211-10K11, *** SEQUENCING IN P

HTG

29662: BQ783369
fab29e04.y1 Zebrafish SJD 15 day embryo Danio rerio cDNA clone IMAGE:5904654 5', MRNA sequence
gi|21991841|gb|BQ783369.1|[21991841]

EST

Bulk GenBank Records

29663: BQ783368
fab29a03.y1 Zebrafish SJD 15 day embryo Danio rerio cDNA clone IMAGE:5904652 5' similar to TR:Q9QXP0 Q9QXP0 RH TYPE C GLYCOPROTEIN. .; MRNA sequence
gi|21991840|gb|BQ783368.1|[21991840]

EST

29664: BQ783367
fab29a01.y1 Zebrafish SJD 15 day embryo Danio rerio cDNA clone IMAGE:5904648 5' similar to SW:COXG_BOVIN P00429 CYTOCHROME C OXIDASE POLYPEPTIDE VIB .; MRNA sequence
gi|21991839|gb|BQ783367.1|[21991839]

EST

29665: AF434664
Danio rerio insulin-like growth factor binding protein-1 mRNA, complete cds
gi|21954646|gb|AF434664.1|[21954646]

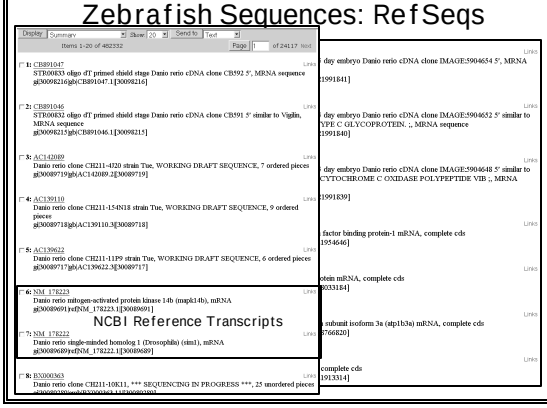
29666: AF330001
Danio rerio UNC45-related protein mRNA, complete cds
gi|18033184|gb|AF330001.1|[18033184]

29667: AF469651
Danio rerio Na,K-ATPase beta subunit isoform 3a (atp1b3a) mRNA, complete cds
gi|18766820|gb|AF469651.1|[18766820]

29668: AF527755
Danio rerio Cyp11a1 mRNA, complete cds
gi|21913314|gb|AF527755.1|[21913314]

NCBI

Zebrafish Sequences: RefSeqs



NCBI

Molecular Databases

- Primary
  - Data provided by experimentalist
  - Record maintained by submitter
  - GenBank
- Derivative
  - Value added to primary data
    - compilation
    - curation
    - assembly
  - Record maintained by database staff
  - NCBI Reference Sequences (RefSeq)
  - Protein Data

NCBI

GenBank: NCBI's Primary Sequence Database

Release 135	April 2003
24,027,936	Records
31,099,264,455	Nucleotides
120,000 +	Species

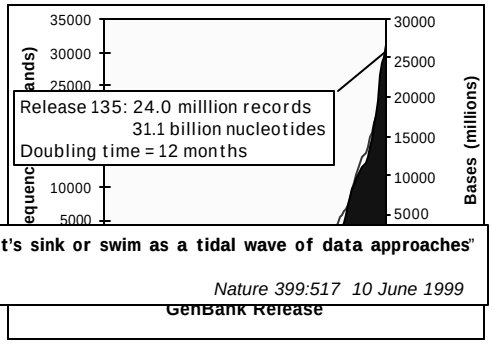
- full release every two months
- incremental and cumulative updates daily
- available only through internet

ftp://ftp.ncbi.nih.gov/genbank/

114 Gigabytes

NCBI

The Growth of GenBank



"It's sink or swim as a tidal wave of data approaches"

Nature 399:517 10 June 1999

NCBI

Traditional GenBank Divisions

- Direct Submissions (Sequin and BankIt)
- Accurate
- Well characterized

BCT	Bacterial and Archeal
INV	Invertebrate
MAM	Mammalian (ex. ROD and PRI)
PHG	Phage
PLN	Plant and Fungal
PRI	Primate
ROD	Rodent
SYN	Synthetic (vectors, synth. genes)
VRL	Viral
VRT	Other Vertebrate

NCBI

Traditional GenBank Records

```
LOCUS       AF105152                1584 bp    mRNA    linear   VRT 11-FEB-1999
DEFINITION  Danio rerio rhodopsin mRNA, complete cds.
ACCESSION   AF105152
VERSION     AF105152.1  GI:4262520
KEYWORDS    .
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
REFERENCE
  AUTHORS
  TITLE
  JOURNAL
  MEDLINE
   PUBMED
REFERENCE
  AUTHORS
  TITLE
  JOURNAL
```

```
LOCUS       AF109368                1581 bp    mRNA    linear   VRT 09-AUG-2001
DEFINITION  Danio rerio rhodopsin mRNA, complete cds.
ACCESSION   AF109368
VERSION     AF109368.1  GI:4581736
KEYWORDS    .
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
            Cypriniformes; Cyprinidae; Danio.
REFERENCE   1  (bases 1 to 1581)
  AUTHORS   Vihtelic,T.S., Doro,C.J. and Hyde,D.R.
  TITLE     Cloning and characterization of six zebrafish photoreceptor opsin
            cDNAs and immunolocalization of their corresponding proteins
  JOURNAL   Vis. Neurosci. 16 (3), 571-585 (1999)
  MEDLINE   99277479
   PUBMED   10349976
REFERENCE   2  (bases 1 to 1581)
  AUTHORS   Vihtelic,T.S., Doro,C.J. and Hyde,D.R.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-NOV-1998) Biological Sciences, University of Notre
            Dame, 264 Galvin Life Science Bldg., Notre Dame, IN 46556, USA
```

NCBI

Traditional GenBank Records

```
FEATURES             Location/Qualifiers
     source          1..1581
                     /organism="Danio rerio"
                     /mol_type="mRNA"
                     /db_xref="taxon:7955"
     CDS             102..1166
                     /note="seven transmembrane light receptor; immunolocalized
                     to the rod outer segments"
                     /codon_start=1
                     /product="rhodopsin"
                     /protein_id="AAD24751.1"
                     /db_xref="GI:4581737"
                     /translation="MNGTEGPAFYVPMSNATGVVRSPYEYPQYYLVAPWAYGLLAAYM
                     FFLIITGFPVNFLTLYVTIEHKKLRTPLNYILLNLAIADLFMVFGGFTTTMYTSLHGY
                     FVFGRLGCNLEGFFATLGGEMGLWSLVVLAIERWMVVCKPVSNFRFGENHAIMGVAFT
                     WVMACSCAVPPLVGWSRYIPEGMQCSCGVDYYTRTPGVNNESFVIYMFIVHFIIPLIV
                     IFFCYGRLVCTVKEAAAQQQESETTQRAEREVTRMVIIMVIAFLICWLPYAGVAWYIF
                     THGGSEFGPVFMTLPAFFAKTSAVYNPCIYICMNKQFRHCMITTLCCGKNPFEEEEGA
                     STTASKTEASSVSSSSVSPA"
```

NCBI

Traditional GenBank Records

```
FEATURES             Location/Qualifiers
     source          1..1581
     CDS
```

```
BASE COUNT      344 a    503 c    383 g    351 t
ORIGIN
        1 gaattcggca cgagcacggt cctgcctcgt ttcttcacag tcctgcccag acatctagag
       61 attaccgcag aaggggctga gcgccacatc caaccgcagc catgaacggt acagagggac
      121 cggcattcta cgtgcctatg tccaatgcca ccggggtagt ccggagcccca tacgaatacc
      181 cacagtacta cctggtggcg ccctgggcat atggcttgct ggcggcctac atgttcttcc
      241 tcatcatcac cggctcccc gtcaacttcc tcacgctgta cgtcaccatc gagcacaaga
      301 agctgcgcac accccctcaac tacatcctgc tgaacctggc cattgccgac ctcttcatgg
      361 tgttcggcgg cttcaccacc accatgtaca cctccttgca cggctacttc gtgttcggcc
      421 gcctgggctg caacctggag ggcttctttg ccaccctggg cggtgagatg gggctatggt
      481 cgcttgtagt actggccatt gagcgctgga tgtgtgtctg caaaccggtc agcaacttcc
      541 gtttcgggga gaaccatgcc atcatggggg tggccttcac ctgggtcatg gcctgctcct
      601 gtgccgtgcc gccactggtg ggctggtcgc gttacatccc cgagggcatg cagtgctcct
      661 gcggcgtcga ctactacacc cgcactcccg gtgtcaacaa cgagtccttc gtcatctaca
      721 tgttcatcgt gcacttcttc atcccactga tagtcatctt cttctgctat ggccggctgg
      781 tgtgcactgt gaaggaggcg gcagcgcagc agcaggagtc cgagaccaca cagcgggcog
      841 agcgcgaggt cacccgcatg gtcatcatca tggtcatcgc cttcctcatc tgctggttgc
      901 cctacgccgg cgtggcctgg tacatcttca cccaccaggg cagcgagttc gggcccgtct
      961 tcatgacccct gcccgcttcc ttcgccaaga cgtccgccgt ctacaacccc tgcatctaca
     1021 tctgcatgaa caagcagttc cgtcactgca tgatcaccac cctgtgctgc ggcaagaacc
     1081 ccttcgaaga ggaagagggc gcctccacca ccgcctccaa gaccgaggct tcgtccgtgt
     1141 cttccagctc cgtgtctccg gcgtaaactc tccttccaaa cactctcaaa gcaccaccct
     1201 gcactgggct cccacggcaa accacacaga g
     1261 gcgaccgcta ccacttgcag aaaaaaaatg t
     1321 aaacccaatt ggttcaacca aaagacagtt t
     1381 agtacatcat ccagcgagtc cagcgtaacg g
     1441 taaaacgcag cagaaggaaa cacatcttaa c
     1501 ttgttgtgta ttgtagaggc atgtattcaa ggcaacgtaa caataaaaag cactttgcaa
     1561 attaaaaaaa aaaaaaaaaa a
//
```

gbdiv_vrt[Properties]
zebrafish: 3,816

NCBI

Bulk GenBank Divisions

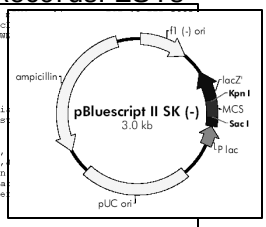
- Batch Submission and htg (email and ftp)
- Inaccurate
- Poorly Characterized

EST	Expressed Sequence Tag
STS	Sequence Tagged Site
GSS	Genome Survey Sequence
HTG	High Throughput Genomic

NCBI

Bulk GenBank Records: ESTs

```
LOCUS       BG738754                 530 bp
DEFINITION  fp62d08.yl Zebrafish adult retina c
            IMAGE:4786599 5' similar to TR:Q9PWN
            sequence.
ACCESSION   BG738754
VERSION     BG738754.1  GI:14088443
KEYWORDS    EST.
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Cranio
            Actinopterygii; Neopterygii; Teleos
            ; Cyprinidae; Danio.
REFERENCE   1  (bases 1 to 530)
  AUTHORS   Clark,M., Johnson,S.L., Lehrach,H.,
            ,S., Hillier,L., Kucaba,T., Martin,
            ,K., Steptoe,M., Theising,B., Allen
            Swaller,T., Gibbons,M., Pape,D., Ha
            Kohn,S., Shin,T., Jackson,Y., Carde
            and Wilson,R.
  TITLE     WashU Zebrafish EST Project 1998
  JOURNAL   Unpublished (1998)
COMMENT     Contact: Stephen L. Johnson
            Washington University School of
            4444 Forest Park Parkway, Box 85
            Tel: 314 286 1800
            Fax: 314 286 1810
            Email: zbrafish@watson.wustl.edu
            Library constructed by: Chandra
            Sequencing by: Washington Univer
            distribution: RessourcenZentrumP
            (web address: www.rzpd.de)
            Seq primer: T3 ET from Amersham
            High quality sequence stop: 423.
```



- first pass, single read cDNA
- largest GenBank division
- gbdiv_est[Properties]
- zebrafish ESTs: 309,433

NCBI

Bulk GenBank Records: ESTs

```
LOCUS       BG738754                 530 bp    mRNA    linear   EST 13-FEB-2002
DEFINITION  fp62d08.yl Zebrafish adult retina cDNA Danio rerio cDNA clone
            IMAGE:4786599 5' similar to TR:Q9PWN4 Q9PWN4 RHODOPSIN. ;, mRNA
            sequence.
ACCESSION   BG738754
VERSION     BG738754.1  GI:14088443
KEYWORDS    EST.
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
            ; Cyprinidae; Danio.
REFERENCE   1  (bases 1 to 530)
  AUTHORS   Clark,M., Johnson,S.L., Lehrach,H., Hillier,L., Kucaba,T.,
            ,S., Hillier,L., Kucaba,T., Martin,K., Steptoe,M., Theising,B.,
            ,K., Steptoe,M., Theising,B., Allen,M., Dodge,I., Pape,D.,
            Swaller,T., Gibbons,M., Pape,D., Hall,L., Wilson,R.,
            Kohn,S., Shin,T., Jackson,Y., Carde
            and Wilson,R.
  TITLE     WashU Zebrafish EST Project 1998
  JOURNAL   Unpublished (1998)
COMMENT     Contact: Stephen L. Johnson
            Washington University School of
            4444 Forest Park Parkway, Box 85
            Tel: 314 286 1800
            Fax: 314 286 1810
            Email: zbrafish@watson.wustl.edu
            Library constructed by: Chandra
            Sequencing by: Washington Univer
            distribution: RessourcenZentrumP
            (web address: www.rzpd.de)
            Seq primer: T3 ET from Amersham
            High quality sequence stop: 423.
FEATURES             Location/Qualifiers
     source          1..530
                     /organism="Danio rerio"
                     /strain="wild-type"
                     /db_xref="taxon:7955"
                     /clone="IMAGE:4786599"
                     /clone_lib="Zebrafish adult retina cDNA"
                     /sex="mixed"
                     /dev_stage="1-2 years"
                     /lab_host="E.Coli XL1-Blue MRF' (XL1-Blue MRF')"
                     /note="Vector: Lambda ZAP II (pBluescript SK-); Site_1:
                     EcoRI; Site_2: SalI; This Zebrafish library was
                     constructed by Dr. Susan E. Brockerhoff (email:
                     sbrocker@u.washington.edu) RZPD library number: 760"
BASE COUNT      101 a    176 c    136 g    117 t
ORIGIN
        1 ttcttcacag tcctgcccag acatctagag attaccgcag aaggggctga gcgccacatc
       61 caaccgcagc catgaacggt acagagggac cggcattcta cgtgcctatg tccaatgcca
      121 ccgggtagt ccggagcccca tacgaatacc cacagtacta cctggtggcg ccctgggcat
      181 atggcttgct ggcggcctac atgttcttcc tcatcatcac cggcttcccc gtcaacttcc
      241 tcacgctgta cgtcaccatc gagcacaaga agctgcgcac acccctcaac tacatcctgc
      301 tgaacctggc cattgccgac ctcttcatgg tgttcggcgg cttcaccacc accatgtaca
      361 cctccttgca cggctacttt gtgttcggcc gcctgggctg caacctggag ggcttctttg
      421 ccaccctggg cggtgagatg ggcctatggt cgcttgcagt actggccatt gagcgctgga
      481 tggtggtctg caaaccggtc caccacttcc gtttcgggga gaaccatgcc
//
```

NCBI

What is UniGene?

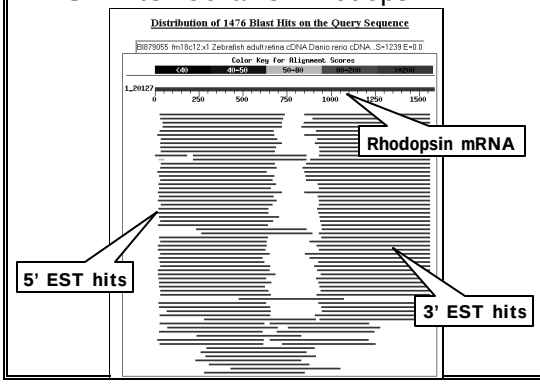
A gene-oriented view of sequence entries

- **MegaBlast based automated sequence clustering**
- **Nonredundant set of gene oriented clusters**
- **Each cluster a unique gene**
- **Information on tissue types and map locations**
- **Includes well-characterized genes and novel ESTs**
- **Useful for gene discovery and selection of mapping reagents**

http://www.ncbi.nlm.nih.gov/UniGene/

NCBI

EST hits zebrafish rhodopsin mRNA



NCBI

UniGene Collections: Animals

		Sequences	Clusters
Homo sapiens	human	4,135,222	111,064
Mus musculus	mouse	3,442,382	90,444
Rattus norvegicus	rat	346,154	63,253
Bos taurus	cow	130,603	10,675
Sus scrofa	pig	62,004	14,390
Gallus gallus	chicken	125,827	5,042
Danio rerio	zebrafish	217,019	15,707
Oryzias latipes	medaka	81,296	6,115
Xenopus laevis	frog	177,814	19,045
Siluriana tropicalis	frog	110,976	7,674
Ciona intestinalis	tunicate	464,060	13,328
Drosophila melanogaster	fruit fly	276,922	14,701
Anopheles gambiae	mosquito	91,352	3,243
Caenorhabditis elegans	nematode	200,436	19,045

March 2003

NCBI

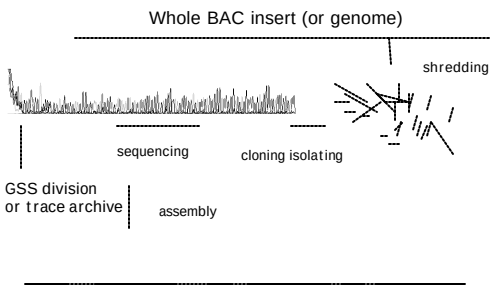
UniGene Collections: Plants and Others

		Sequences	Clusters
Arabidopsis thaliana	thale cress	221,113	26,813
Glycine max	soybean	201,710	9,084
Medicago truncatula	barrel medic	97,063	5,729
Lycopersicon esculentum	tomato	70,118	3,749
Triticum aestivum	wheat	297,961	20,376
Hordeum vulgare	barley	263,407	10,272
Zea mays	maize (corn)	163,541	13,512
Oryzia sativa	rice	94,780	19,191
Sorghum bicolor	sorghum	17,742	802
Chlamydamonas reinhardti	alga	80,771	5,712
Dictyostelium discoideum	slime mold	119,140	4,782

March 2003

NCBI

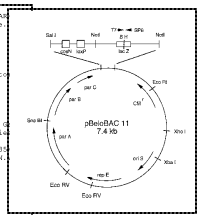
Genome Sequencing - HTG, GSS,(WGS)



NCBI

Bulk GenBank Records: GSS

```
LOCUS       BX247846                 551 bp    DNA     linear   GSS 14-MAR
DEFINITION  Danio rerio genomic clone DKEY-235N9, genomic survey sequence.
ACCESSION   BX247846
VERSION     BX247846.1  GI:28170522
KEYWORDS    GSS.
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
            Cypriniformes; Cyprinidae; Danio.
REFERENCE   1  (bases 1 to 551)
  AUTHORS   Humphray,S.J., Huckle,E. and Durham,J.L.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-MAR-2003) The Sanger Institute, Wellcome Trust G
            Campus, Hinxton, Cambridgeshire, CB10 1SA, UK. E-mail enquiries
            humquery@sanger.ac.uk Unpublished
COMMENT     This sequence was generated from the SP6 end of BAC 235N9. 235
            part of the Daniokey BAC Library created by R. Plasterk and N.
            Keygene. Further details:
            http://www.sanger.ac.uk/Projects/D_rerio/.
FEATURES             Location/Qualifiers
     source          1..551
                     /organism="Danio rerio"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:7955"
                     /clone="DKEY-235N9"
                     /tissue_type="Testis"
                     /note="vector pIndigoBAC-536"
BASE COUNT      205 a    71 c   110 g   165 t
ORIGIN
        1 ctttatttaa attttctact gatgaaaaca aaacacgcaa t
       61 aaactaagtc aattaatagg tgaagagggt ttataaatga c
      121 acttgcagac tggaaatgtg gagaaactgg aaataccacca g
      181 gtctggtgtt ctgttgcagg aaattaagac aaaacaggag a
      241 ggatcattaa ttaacacatc agtgagatca acaatcaagt a
      301 atagdttgga ataaggagaa aggggttgtt ggtgggaata g
      361 aactgagtaa tttctgttgc atcaaccttt taatttttcca c
      421 aaactgcatc atgactcaga ctttgatatg aaggataaga ggacattttt aatatttgtg
      481 tgatgacaga aaattcagaa attcattaaa acgaacataa gaaatatacc acgacatgtc
      541 agcgtataaa t
//
```



- first pass, single read gDNA
- surveys of BAC libraries
- gbdiv_gss[Properties]
- zebrafish GSS: 159, 024

NCBI

HTG Division: High Throughput Genome

Zebrafish PAC Clone

phase 1 → ← → → ← HTG
Acc = AC109580.1

phase 2 → → → HTG
Acc =AC109580.11

phase 3 → VRT
Acc = AC109580.14

gbdiv_htg[Properties]
zebrafish: 2,958

40,000 to > 350,000 bp

NCBI

HTG Division: High Throughput Genome

```
LOCUS       AC109580                2463 bp    DNA     linear   HTG 05-FEB-2002
DEFINITION  Danio rerio clone busm1-183j13 strain AB, WORKING DRAFT SEQUENCE, 1
            unordered piece.
ACCESSION   AC109580
VERSION     AC109580.1  GI:18497096
KEYWORDS    HTG; HTGS_PHASE1; HTGS_DRAFT.
SOURCE      Danio rerio (zebrafish)
```

```
LOCUS       AC109580               97770 bp    DNA     linear   HTG 26-OCT-2002
DEFINITION  Danio rerio clone busm1-183j13 strain AB, WORKING DRAFT SEQUENCE, 4
            ordered pieces.
ACCESSION   AC109580
VERSION     AC109580.11  GI:24414284
KEYWORDS    HTG; HTGS_PHASE2; HTGS_DRAFT.
SOURCE      Danio rerio (zebrafish)
```

```
LOCUS       AC109580               94348 bp    DNA     linear   VRT 19-NOV-2002
DEFINITION  Danio rerio clone busm1-183j13 strain AB, complete sequence.
ACCESSION   AC109580
VERSION     AC109580.14  GI:24635954
KEYWORDS    HTG.
SOURCE      Danio rerio (zebrafish)
```

40,000 to > 350,000 bp

NCBI

Whole Genome Shotgun

NCBI

Whole Genome Shotgun in GenBank

1: AABT00000000
Aspergillus terreus, whole genome shotgun sequencing project
gi|27262064|gb|AABT00000000.1|AABT01000000[27262064]

2: AABU00000000
Drosophila melanogaster strain y; cn b
sequencing project
gi|27232097|gb|AABU00000000.1|AAB

3: AABS00000000
Ciona intestinalis, whole genome shotg
gi|26556995|gb|AABS00000000.1|AAB

4: AABR00000000
Rattus norvegicus strain BN/SsNHsd/M
sequencing project
gi|25762278|gb|AABR00000000.1|AAB

5: AABL00000000
Plasmodium yoelii yoelii, whole genom
gi|23491527|gb|AABL00000000.1|AAB

6: CAAB00000000
Takifugu rubripes, whole genome shot
gi|22418063|emb|CAAB00000000.1|CA

wgs_master[Properties]

7: CAAC00000000
Caenorhabditis briggsae, whole genome shotgun sequencing project
gi|22417469|emb|CAAC00000000.1|CAAC01000000[22417469]

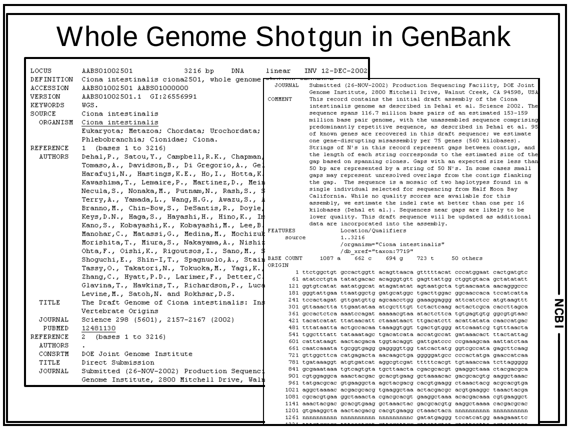
8: AAAB00000000
Anopheles gambiae str. PEST whole genome shotgun sequencing project
gi|19683831|gb|AAAB00000000.1|AAAB01000000[19683831]

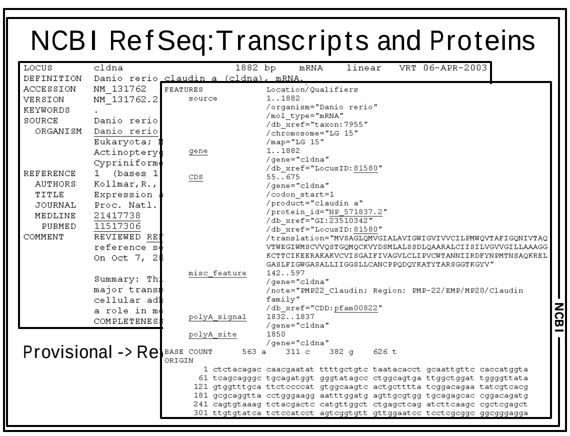
9: AAAD00000000
Mus musculus chromosome 16, whole genome shotgun sequencing project
gi|21281728|gb|AAAD00000000.1|AAAD01000000[21281728]

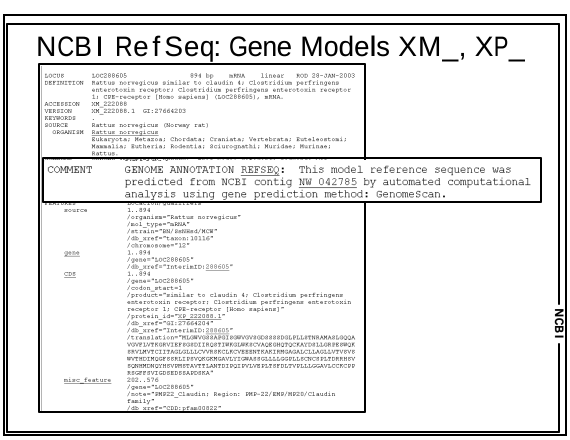
10: CAAA00000000
Mus musculus, whole genome shotgun sequencing project
gi|20800445|emb|CAAA00000000.1|CAAA01000000[20800445]

11: AAAA00000000
Oryza sativa (indica cultivar-group) cultivar 93-11 whole genome shotgun sequencing project
gi|19924305|gb|AAAA00000000.1|AAAA01000000[19924305]

NCBI


Whole Genome Shotgun in GenBank
LOCUS AABS01002501 3216 bp DNA linear INV 12-DEC-2002
DEFINITION Ciona intestinalis ciona2501, whole genome
ACCESSION AABS01002501 AABS01000000
VERSION AABS01002501.1 GI:26556991
KEYWORDS WGS.
SOURCE Ciona intestinalis
ORGANISM Ciona intestinalis
Eukaryota; Metazoa; Chordata; Urochordata;
Phlebobranchia; Cionidae; Ciona.
REFERENCE 1 (bases 1 to 3216)
AUTHORS Dehal,P., Satou,Y., Campbell,R.K., Chapman,
Tomaso,A., Davidson,B., Di Gregorio,A., Ge
Harafuji,N., Hastings,K.E., Ho,I., Hotta,K.
Kawashima,T., Lemaire,P., Martinez,D., Mei
Necula,S., Nonaka,M., Putnam,N., Rash,S.,
Terry,A., Yamada,L., Wang,H.G., Awazu,S.,
Branno,M., Chin-Bow,S., DeSantis,R., Doyle,
Keys,D.N., Haga,S., Hayashi,H., Hino,K., I
Kano,S., Kobayashi,K., Kobayashi,M., Lee,B.
Manohar,C., Matassi,G., Medina,M., Mochizu
Morishita,T., Miura,S., Nakayama,A., Nishi
Ohta,F., Oishi,K., Rigoutsos,I., Sano,M.,
Shoguchi,E., Shin-I,T., Spagnuolo,A., Stai
Tassy,O., Takatori,N., Tokuoka,M., Yagi,K.
Zhang,C., Hyatt,P.D., Larimer,F., Detter,C.
Glavina,T., Hawkins,T., Richardson,P., Luca
Levine,M., Satoh,N. and Rokhsar,D.S.
TITLE The Draft Genome of Ciona intestinalis: Ins
Vertebrate Origins
JOURNAL Science 298 (5601), 2157-2167 (2002)
PUBMED 12481130
REFERENCE 2 (bases 1 to 3216)
AUTHORS .
CONSRTM DOE Joint Genome Institute
TITLE Direct Submission
JOURNAL Submitted (26-NOV-2002) Production Sequenc
Genome Institute, 2800 Mitchell Drive, Waln
JOURNAL Submitted (26-NOV-2002) Production Sequencing Facility, DOE Joint
Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
COMMENT This record contains the initial draft assembly of the Ciona
intestinalis genome as described in Dehal et al. Science 2002. The
sequence spans 116.7 million base pairs of an estimated 153-159
million base pair genome, with the unassembled sequence comprising
predominantly repetitive sequence, as described in Dehal et al. 95
of known genes are recovered in this draft sequence; we estimate
one gene-disrupting misassembly per 75 genes (560 kilobases).
Strings of N's in this record represent gaps between contigs, and
the length of each string corresponds to the estimated size of the
gap based on spanning clones. Gaps with an expected size less than
50 bp are represented by a string of 50 N's. In some cases small
gaps may represent unresolved overlaps from the contigs flanking
the gap. The sequence is a mosaic of two haplotypes found in a
single individual selected for sequencing from Half Moon Bay
California. While no quality scores are available for this
assembly, we estimate the indel rate at better than one per 16
kilobases (Dehal et al.). Sequences near gaps are likely to be
lower quality. This draft sequence will be updated as additional
data are incorporated into the assembly.
FEATURES Location/Qualifiers
source 1..3216
/organism="Ciona intestinalis"
/db_xref="taxon:7719"
BASE COUNT 1087 a 662 c 694 g 723 t 50 others
ORIGIN
1 ttctggctgt gccactggtt acagttaaca gtttttacat cccatggaat cactgatgtc
61 atatcctgta tatatgacac acagggtgtt gagttattgg ctggtgtaca gctatatatt
121 ggtgtcatat aatatggcat atagatatat agtaatgcta tgtaacaata aacagggccc
181 gggtattgaa ttaatggctg gatgcatggc tgacttggac ggcaaccaca tccatcatta
241 tccactagat gttgatgttg agcaacctgg gaaaggaggg attcatctcc atgtaagttt
301 gttaaactta ttgaatataa atcgctttgt tctactcaag actactcgca caccttagca
361 gccactctca aaatccagat aaaaacgtaa atactcttca tgtgagtgtg ggcgtgtaac
421 tacatcatat ttataacatt ctaaataact ttgacatctt acattatata caaccatgac
481 tttataatta actgccacaa taaaggtggt tgactgtggg attcaaatcg tgtttaacta
541 tggctttatt tataaatagc tgacatcata accatgccat gataaacact ttactattag
601 cattataagt aactacgaca tggtacaggt gattgatccc ccgaaagcaa aattatctaa
NCBI


NCBI RefSeq:Transcripts and Proteins
LOCUS cldna 1882 bp mRNA linear VRT 06-APR-2003
DEFINITION Danio rerio claudin a (cldna), mRNA.
ACCESSION NM_131762
VERSION NM_131762.2
KEYWORDS .
SOURCE Danio rerio
ORGANISM Danio rerio
Eukaryota;
Actinoptery
Cypriniform
REFERENCE 1 (bases 1
AUTHORS Kollmar,R.,
TITLE Expression
JOURNAL Proc. Natl.
MEDLINE 21417738
PUBMED 11517306
COMMENT REVIEWED RE
reference s
On Oct 7, 2
Summary: Th
major trans
cellular ad
a role in m
COMPLETENES
FEATURES Location/Qualifiers
source 1..1882
/organism="Danio rerio"
/mol_type="mRNA"
/db_xref="taxon:7955"
/chromosome="LG 15"
/map="LG 15"
gene 1..1882
/gene="cldna"
/db_xref="LocusID:81580"
CDS 55..675
/gene="cldna"
/codon_start=1
/product="claudin a"
/protein_id="NP_571837.2"
/db_xref="GI:23510342"
/db_xref="LocusID:81580"
/translation="MVSAGLQMVGIALAVIGWIGVIVVCILPMWQVTAFIGQNIVTAQ
VTWEGIWMSCVVQSTGQMQCKVYDSMLALSSDLQAARALCIISILVGVVGILLAAAGG
KCTTCIKEERAKAKVCVISGAIFIVAGVLCLIPVCWTANNIIRDFYNPMTNSAQKREL
GASLFIGWGASALLIIGGSLLCANCPPQDQYKATYTARSGGTKGYV"
misc_feature 142..597
/gene="cldna"
/note="PMP22_Claudin; Region: PMP-22/EMP/MP20/Claudin
family"
/db_xref="CDD:pfam00822"
polyA_signal 1832..1837
/gene="cldna"
polyA_site 1850
/gene="cldna"
BASE COUNT 563 a 311 c 382 g 626 t
ORIGIN
1 ctctacagac caacgaatat ttttgctgtc taatacacct gcaattgttc caccatggta
61 tcagcagggc tgcagatggt gggtatagcc ctggcagtga ttggctggat tggggttata
121 gtggtttgca ttctccccat gtggcaagtc actgctttta tcggacagaa tatcgtcacg
181 gcgcaggtta cctgggaagg aatttggatg agttgcgtgg tgcagagcac cggacagatg
241 cagtgtaaag tctacgactc catgttggct ctgagctcag atcttcaagc cgctcgagct
301 ttgtgtatca tctccatcct agtcggtgtt gttggaatcc tcctcgcggc ggcgggagga
Provisional -> Re
NCBI


NCBI RefSeq: Gene Models XM_, XP_
LOCUS LOC288605 894 bp mRNA linear ROD 28-JAN-2003
DEFINITION Rattus norvegicus similar to claudin 4; Clostridium perfringens
enterotoxin receptor; Clostridium perfringens enterotoxin receptor
1; CPE-receptor [Homo sapiens] (LOC288605), mRNA.
ACCESSION XM_222088
VERSION XM_222088.1 GI:27664203
KEYWORDS .
SOURCE Rattus norvegicus (Norway rat)
ORGANISM Rattus norvegicus
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae;
Rattus.
COMMENT GENOME ANNOTATION REFSEQ: This model reference sequence was
predicted from NCBI contig NW 042785 by automated computational
analysis using gene prediction method: GenomeScan.
source 1..894
/organism="Rattus norvegicus"
/mol_type="mRNA"
/strain="BN/SsNHsd/MCW"
/db_xref="taxon:10116"
/chromosome="12"
gene 1..894
/gene="LOC288605"
/db_xref="InterimID:288605"
CDS 1..894
/gene="LOC288605"
NCBI

RefSeq Contigs: NT_, NW_

```
LOCUS       NT_007933               64383103 bp    DNA     1
DEFINITION  Homo sapiens chromosome 7 genomic contig.
ACCESSION   NT_007933
VERSION     NT_007933.12  GI:29798863
KEYWORDS    .
SOURCE      Homo sapiens
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vert
            Mammalia; Eutheria; Primates; Catarrhini; Ho
REFERENCE   1  (bases 1 to 64383103)
  AUTHORS   International Human Genome Sequencing Consor
```

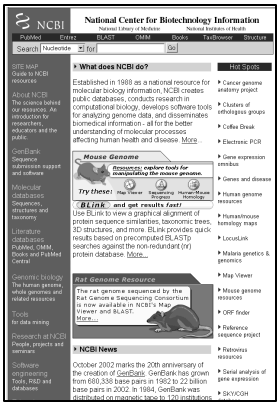
On Apr 11, 2003 this sequen
This contig is part of the
produced by the Human Genom
from its component clone se
provided by the sequencing
contig are those produced
33.

```
     source          1..64383103
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="7"
     STS             10802..10871
                     /standard_name="GDB:4585211"
                     /note="WARNING: map location ambigu
     STS             10910..11046
                     /standard_name="SHGC-56087"
                     /note="WARNING: map location ambigu
                     /db_xref="UniSTS:82487"
     STS             24205..24324
                     /standard_name="RH93206"
                     /note="WARNING: map location ambigu
                     /db_xref="UniSTS:19294"
     STS             35098..35246
                     /standard_name="RH12585"
                     /note="WARNING: map location ambigu
                     /db_xref="UniSTS:19333"
     STS             35131..35233
                     /standard_name="AFM291WG"
```

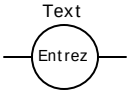
```
LOCUS       NW_042785               1855294 bp    DNA     linear   CON 29-JAN-2003
DEFINITION  Rattus norvegicus chromosome 12 WGS supercontig.
ACCESSION   NW_042785
VERSION     NW_042785.1  GI:26014190
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae;
            Rattus.
COMMENT     GENOME ANNOTATION REPEPO: NCBI contigs are derived from assembled
            genomic sequence data.
            Also see:
                    Documentation of NCBI's Annotation Process

FEATURES             Location/Qualifiers
     source          1..1855294
                     /organism="Rattus norvegicus"
                     /mol_type="genomic DNA"
                     /strain="BN/SsNHsd/MCW"
                     /db_xref="taxon:10116"
                     /chromosome="12"
     STS             2..154
                     /standard_name="G48114"
                     /db_xref="UniSTS:35224"
     STS             22674..22858
                     /standard_name="AU046599"
                     /db_xref="UniSTS:111299"
     gene            24112..39493
                     /gene="LOC288599"
                     /db_xref="InterimID:288599"
     mRNA            join(24112..24176,33705..33892,34965..35029,35110..35206,
                     38380..38517,38962..39493)
                     /gene="LOC288599"
                     /product="similar to Wbscr1 alternative spliced product
                     [Mus musculus]"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BLAST."
                     /transcript_id="XM_213755.1"
                     /db_xref="GI:27664185"
                     /db_xref="InterimID:288599"
     CDS             join(24112..24176,33705..33892,34965..35029,35110..35206,
                     38380..38517,38962..39101)
                     /gene="LOC288599"
                     /codon_start=1
                     /protein_id="XP_213755.1"
                     /db_xref="GI:27664186"
                     /db_xref="InterimID:288599"
     gene            50310..54664
                     /gene="LOC288600"
```

NCBI

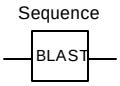
Web Access



Text



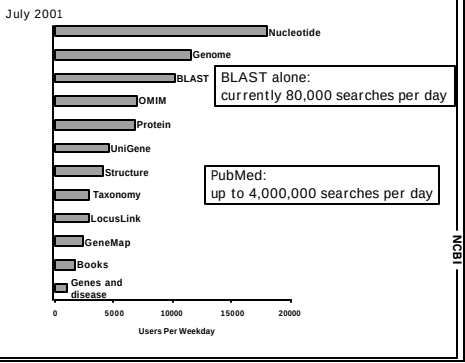
Sequence



Structure

NCBI

Some Web Statistics

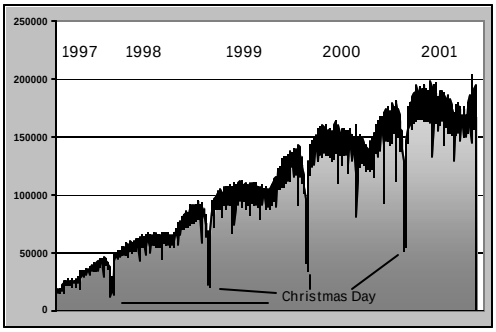


BLAST alone:
currently 80,000 searches per day

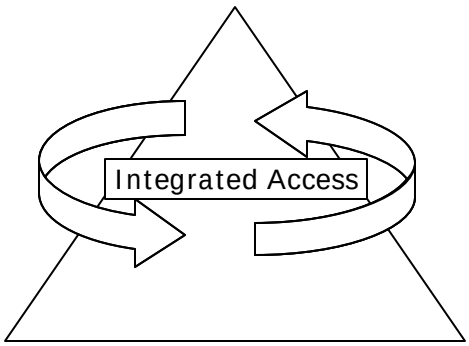
PubMed:
up to 4,000,000 searches per day

NCBI

Users per day

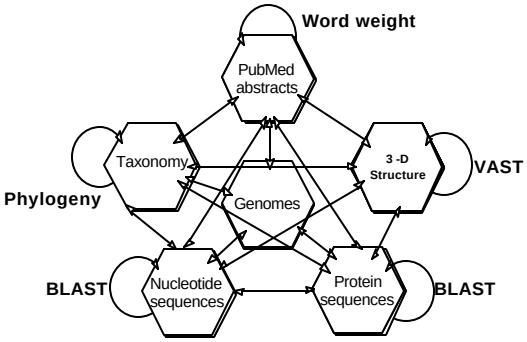


NCBI



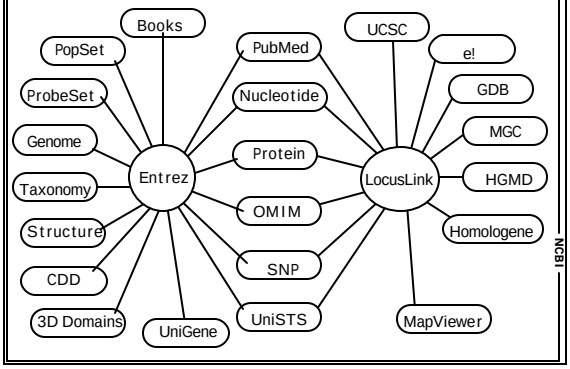
NCBI

Entrez: Database Integration



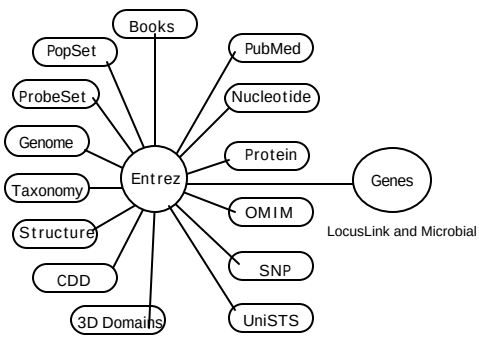
NCBI

The Big Picture



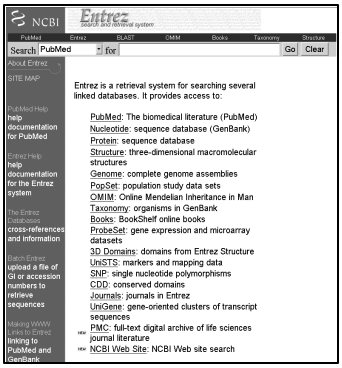
NCBI

The Big Picture



NCBI

WWW Entrez



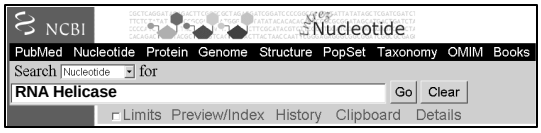
NCBI

Database Searching with Entrez

- Using limits and field restriction to find fish RNA helicases
- Linking and neighboring with RNA helicases

NCBI

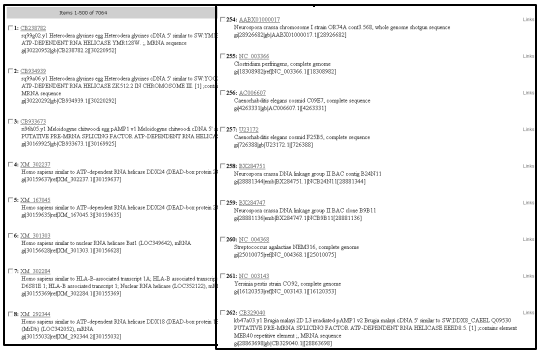
Entrez Nucleotides



NCBI

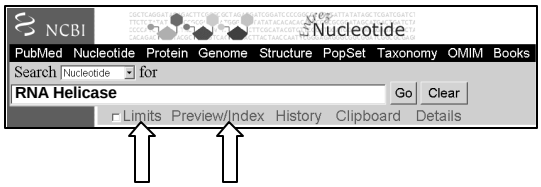
Document Summaries:

RNA Helicase[All Fields]



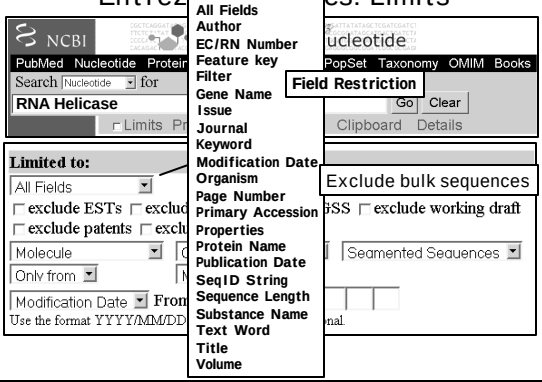
NCBI

Entrez Nucleotides: Limits & Preview/Index



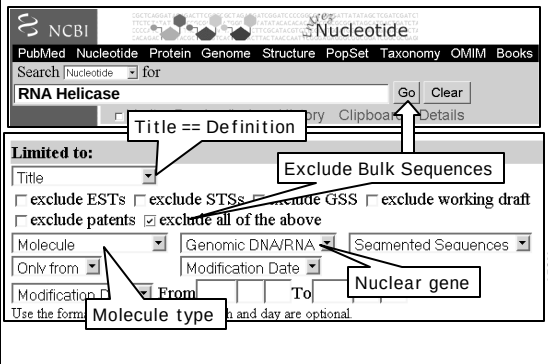
NCBI

Entrez Nucleotides: Limits



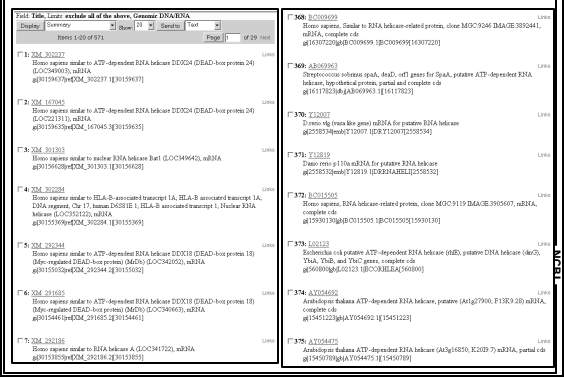
NCBI

Entrez Nucleotides: Limits

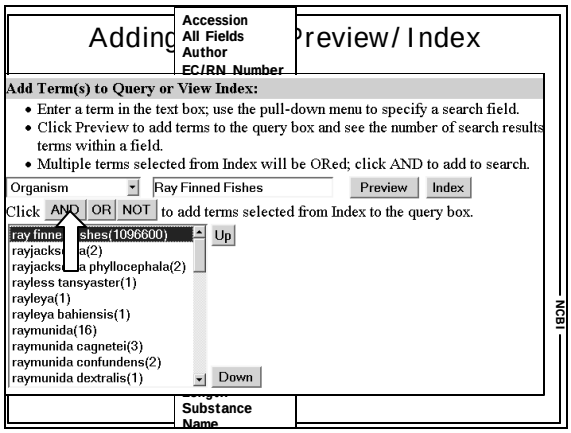


NCBI

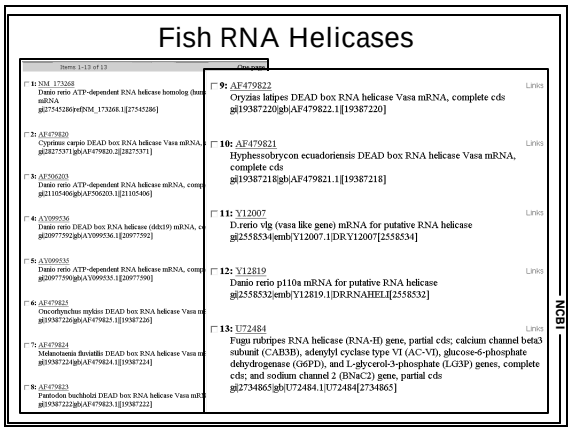
Document Summaries: Limits



Adding Preview/Index



Fish RNA Helicases



Fugu Genomic Region

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Show 20 Send to File Get Subsequence

1: U72484. Fugu rubripes RNA...[gi:2734865]

Links
- Related Sequences
- Protein
- Taxonomy

```
LOCUS       U72484                 61901 bp    DNA     linear   VRT
DEFINITION  Fugu rubripes RNA helicase (RNA-H) gene, partial cds; c
            channel beta3 subunit (CAB3B), adenylyl cyclase type VI (AC-VI),
            glucose-6-phosphate dehydrogenase (G6PD), and
            L-glycerol-3-phosphate (LG3P) genes, complete cds; and sodium
            channel 2 (BNaC2) gene, partial cds.
ACCESSION   U72484
VERSION     U72484.1  GI:2734865
KEYWORDS    .
SOURCE      Takifugu rubripes
  ORGANISM  Takifugu rubripes
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
            Acanthomorpha; Acanthopterygii; Percomorpha; Tetraodontiformes;
            Tetradontoidea; Tetraodontidae; Takifugu.
REFERENCE   1  (bases 1 to 61901)
  AUTHORS   Riboldi Tunnicliffe,G.R., Nyakatura,G., Bauer,D., Elgar,G.S.,
            Perret,X., Brenner,S. and Rosenthal,A.
  TITLE     Sequencing of a 62kb region of Fugu rubripes around the G6PD locus
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 61901)
  AUTHORS   Riboldi Tunnicliffe,G.R., Nyakatura,G., Bauer,D., Perret,X.,
            Elgar,G.S., Brenner,S. and Rosenthal,A.
  TITLE     Direct Submission
```

NCBI

Fugu Genomic Region

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Show 20 Send to File Get Subsequence

1: U72484. Fugu rubripes RNA...[gi:2734865] Links

```
LOCUS       U72484                 61901 bp    DNA     linear   VRT 16-JAN-1998
DEFINITION  Fugu rubripes RNA helicase (RNA-H) gene, partial cds; calcium
            channel beta3 subunit (CAB3B), adenylyl cyclase type VI (AC-VI),
            glucose-6-phosphate dehydrogenase (G6PD), and
            L-glycerol-3-phosphate (LG3P) genes, complete cds; and sodium
            channel 2 (BNaC2) gene, partial cds.
ACCESSION   U72484
VERSION     U72484.1  GI:2734865
KEYWORDS    .
SOURCE      Takifugu rubripes
  ORGANISM  Takifugu rubripes
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
            Acanthomorpha; Acanthopterygii; Percomorpha; Tetraodontiformes;
            Tetradontoidea; Tetraodontidae; Takifugu.
REFERENCE   1  (bases 1 to 61901)
  AUTHORS   Riboldi Tunnicliffe,G.R., Nyakatura,G., Bauer,D., Elgar,G.S.,
            Perret,X., Brenner,S. and Rosenthal,A.
  TITLE     Sequencing of a 62kb region of Fugu rubripes around the G6PD locus
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 61901)
  AUTHORS   Riboldi Tunnicliffe,G.R., Nyakatura,G., Bauer,D., Perret,X.,
            Elgar,G.S., Brenner,S. and Rosenthal,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-OCT-1997) Genome Analysis, Institute of Molecular
```

NCBI

Fugu Genomic Region: Genes

RNA Helicase, partial

```
gene            <1..1564
                /gene="RNA-H"
CDS             join(<5..150,232..383,470..578,689..863,945..1168)
                /gene="RNA-H"
                /codon_start=3
```

calcium channel, beta 3 subunit

```
gene            /gene="RNA-H"
                complement(1705..14167)
                /gene="CAB3B"
polyA_signal    complement(1705..1710)
                /gene="CAB3B"
CDS             complement(join(3496..3888,5400..5570,5650..5745,
                5808..5959,6033..6133,6235..6293,6502..6579,6743..6762,
                6848..6927,7000..7115,7220..7342,8259..8378,12351..12443,
                12905..12985))
```

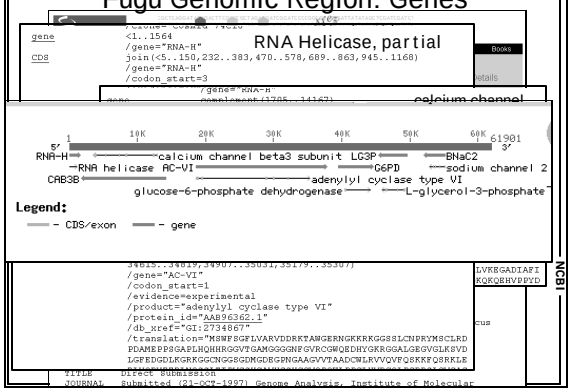
adenylate cyclase 6

```
gene            18610..37777
                /gene="AC-VI"
CDS             join(18610..19139,19226..19541,26886..29035,29974..30095,
                30642..30753,30828..30955,31042..31200,31282..31423,
                31555..31683,31767..31795,32134..32281,32393..32490,
                32766..32853,32932..33048,33202..33360,33466..33641,
                33769..33795,33880..33909,33998..34130,34260..34523,
                34615..34819,34907..35031,35179..35307)
                /gene="AC-VI"
                /codon_start=1
                /evidence=experimental
                /product="adenylyl cyclase type VI"
                /protein_id="AAB96362.1"
                /db_xref="GI:2734867"
                /translation="MSWFSGFLVARVDDRKTAWGERNGKKRKGGSSLCNPRYMSCLRD
                PDAMEPPSGAPLHQHHRGGVTGAMGGGGNFGVRCGWQEDHYGKRGGALGEGVGLKSVD
                LGFEDGDLKGRKGGCNGGSGDMGDEGPNGAAGVVTAADCWLRVVQVFQSKKFQSRKLE
```

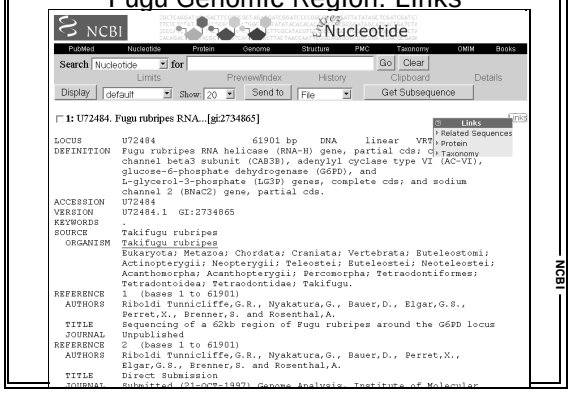
KRKSRFKRSDG
LQLDRAKSKPV
LVKEGADIAFI
KQRQEHVPPVD

NCBI

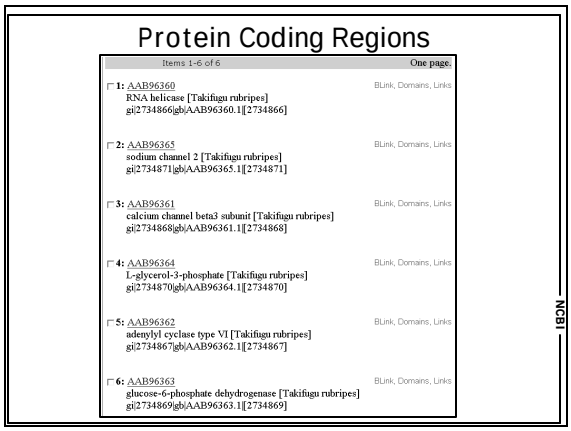
Fugu Genomic Region: Genes



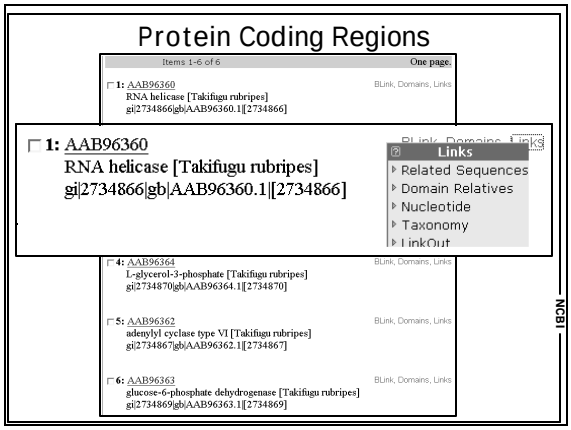
Fugu Genomic Region: Links



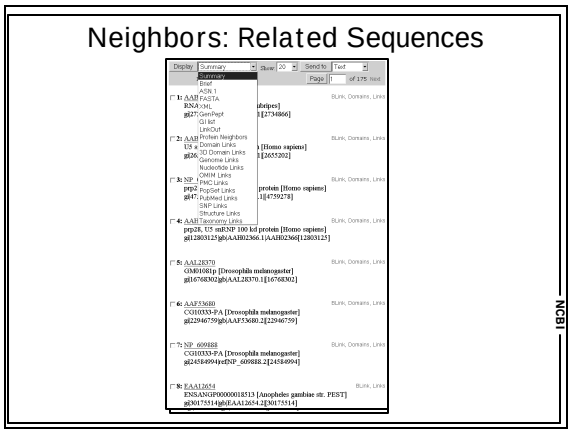
Protein Coding Regions



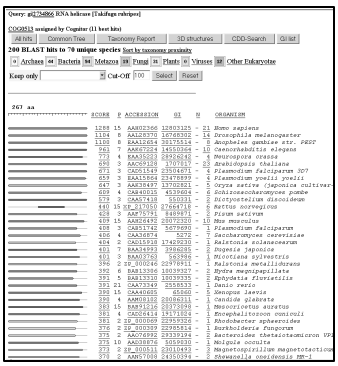
Protein Coding Regions



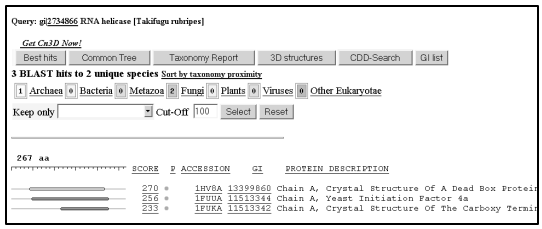
Neighbors: Related Sequences



BLink: Advanced Protein Neighbors

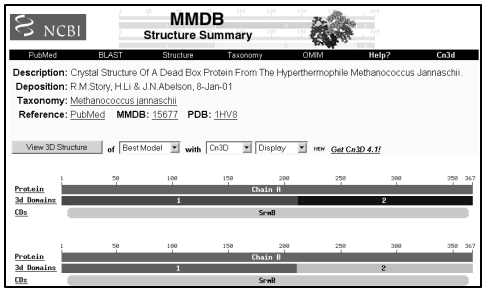


BLink: Structural Models



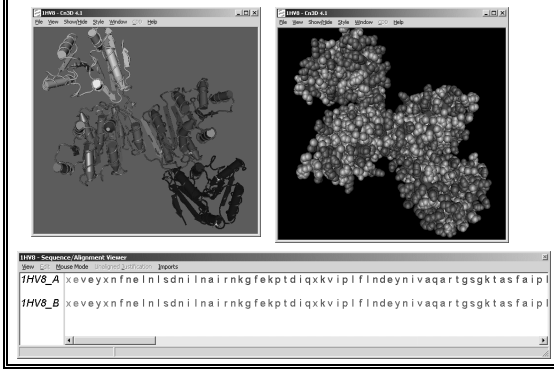
NCBI

Methanococcus DEAD box



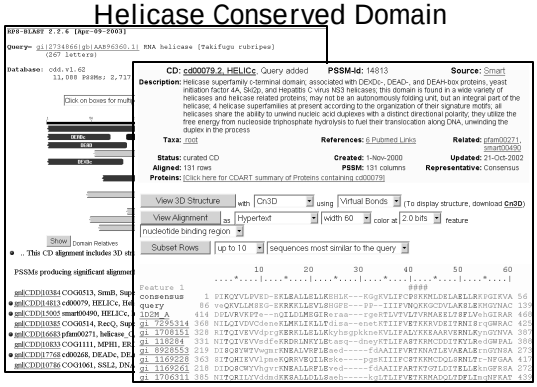
NCBI

Methanococcus DEAD Box: Cn3D

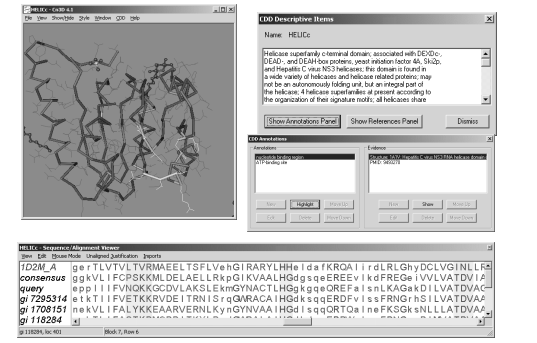


NCBI

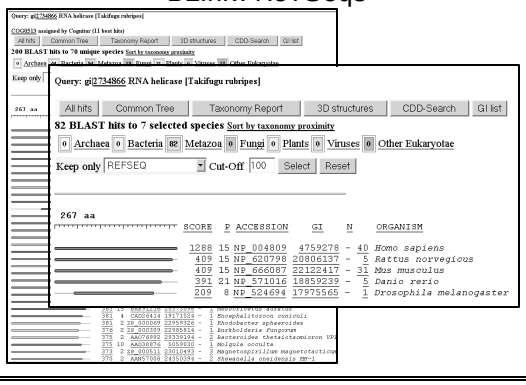
Helicase Conserved Domain



Structural Model for Domain

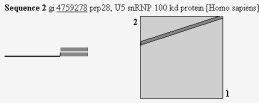


BLink: RefSeqs



Alignment: Fugu vs. Human

Sequence 1 gi 2734866 RNA helicase [Takifugu rubripes] **Length** 267 (1 .. 267)
Sequence 2 gi 4759278 prp28, U5 snRNP 100 kd protein [Homo sapiens] **Length** 820 (1 .. 820)



NOTE The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score =  500 bits (1288), Expect = e-140
Identities = 247/267 (92%), Positives = 258/267 (96%)

Query:                          1   NIDMGFEPDVQKILEFIPVTNQKPDTEEAEDPDKMMNNFESGKNEYRQTVMFTATMPPAV 60
                                    NIDMGFEPDVQKILE +PV+NQKPDT+EAEDP+KM+ NFESGK+EYRQTVMFTATMPPAV
Sbjct:                          554 NIDMGFEPDVQKILEHMPVSNQKPDTDEAEDPEKMLANFESGKHEYRQTVMFTATMPPAV 613
Superfamily II DNA and RNA heli> 554 ************************************************************

Query:                          61  EPLARSYLRRPAVVYIGSAGKPHERVEQKVLLMSEGEKEKKLLKVLSNGFKPPIIIFVNQ 120
                                    EPLARSYLRRPAVVYIGSAGKPHERVEQKV LMSE EEKEKLL +L  GF+PPIIIFVNQ
Sbjct:                          614 EPLARSYLRRPAVVYIGSAGKPHERVEQKVFLMSESEKEKLLALLEQGFPPPIIIFVNQ 673
Superfamily II DNA and RNA heli> 614 ************************************************************

Query:                          121 KKGCDVLAKSLEKMGYNACTLHGGKGQEQREFALSNLKAGAKDILVATDVAGRGIDIHDV 180
                                    KKGCDVLAKSLEKMGYNACTLHGGKGQEQREFALSNLKAGAKDILVATDVAGRGIDI DV
Sbjct:                          674 KKGCDVLAKSLEKMGYNACTLHGGKGQEQREFALSNLKAGAKDILVATDVAGRGIDIQDV 733
Superfamily II DNA and RNA heli> 674 ************************************************************
```

NCBI

Human RefSeq

NCBI Protein

PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM | Books

Search Nucleotide for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display default Show 1 Send to File Get Subsequence

1: NP_004809. prp28, U5 snRNP 1...[gi:4759278]

Links
- Related Sequences
- Domain Relatives
- Map Viewer
- Nucleotide
- PubMed
- SNP
- Taxonomy

```
LOCUS       U5-100K                 820 aa            linear   PR
DEFINITION  prp28, U5 snRNP 100 kd protein [Homo sapiens].
ACCESSION   NP_004809
VERSION     NP_004809.1  GI:4759278
DBSOURCE    REFSEQ: accession NM_004818.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 820)
  AUTHORS   Teigelkamp,S., Mundt,C., Achsel,T., Will,C.L. and Luhrmann,R.
  TITLE     The human U5 snRNP-specific 100-kD protein is an RS
            domain-containing, putative RNA helicase with significant homology
            to the yeast splicing factor Prp28p
  JOURNAL   RNA 3 (11), 1313-1326 (1997)
  MEDLINE   98072280
   PUBMED   9409622
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AF026402.1.
```

NCBI

PubMed Link

1: RNA 1997 Nov;3(11):1313-26

Full text article at www.rnajournal.org

Related Articles | LinkOut

Links
- Nucleotide
- Protein
- Cited in PMC
- Books
- LinkOut

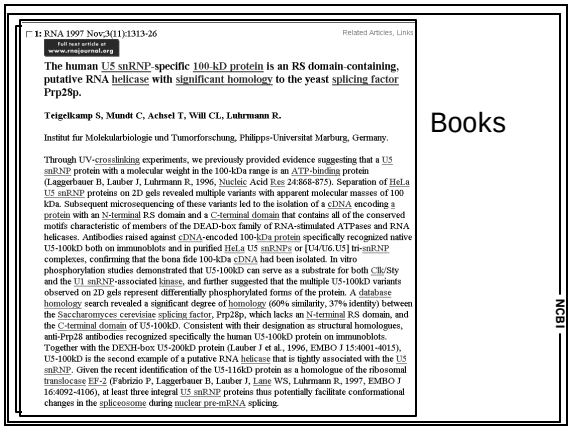
The human U5 snRNP-specific 100-kD protein is an RS domain-c putative RNA helicase with significant homology to the yeast splic Prp28p.

Teigelkamp S, Mundt C, Achsel T, Will CL, Luhrmann R.

Institut fur Molekularbiologie und Tumorforschung, Philipps-Universitat Marburg, Germany.

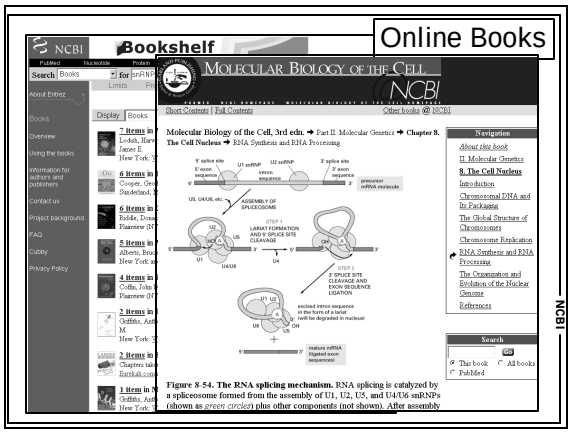
Through UV-crosslinking experiments, we previously provided evidence suggesting that a U5 snRNP protein with a molecular weight in the 100-kDa range is an ATP-binding protein (Laggerbauer B, Lauber J, Luhrmann R, 1996, Nucleic Acid Res 24:868-875). Separation of HeLa U5 snRNP proteins on 2D gels revealed multiple variants with apparent molecular masses of 100 kDa. Subsequent microsequencing of these variants led to the isolation of a cDNA encoding a protein with an N-terminal RS domain and a C-terminal domain that contains all of the conserved motifs characteristic of members of the DEAD-box family of RNA-stimulated ATPases and RNA helicases. Antibodies raised against cDNA-encoded 100-kDa protein specifically recognized native U5-100kD both on immunoblots and in purified HeLa U5 snRNPs or [U4/U6.U5] tri-snRNP complexes, confirming that the bona fide 100-kDa cDNA had been isolated. In vitro phosphorylation studies demonstrated that U5-100kD can serve as a substrate for both Clk/Sty and the U1 snRNP-associated kinase, and further suggested that the multiple U5-100kD variants observed on 2D gels represent differentially phosphorylated forms of the protein. A database homology search revealed a significant degree of homology (60% similarity, 37% identity) between the Saccharomyces cerevisiae splicing factor, Prp28p, which lacks an N-terminal RS domain, and the C-terminal domain of U5-100kD. Consistent with their designation as structural homologues, anti-Prp28 antibodies recognized specifically the human U5-100kD protein on immunoblots. Together with the DEXH-box U5-200kD protein (Lauber J et al., 1996, EMBO J 15:4001-4015), U5-100kD is the second example of a putative RNA helicase that is tightly associated with the U5 snRNP. Given the recent identification of the U5-116kD protein as a homologue of the ribosomal translocase EF-2 (Fabrizio P, Laggerbauer B, Lauber J, Lane WS, Luhrmann R, 1997, EMBO J 16:4092-4106), at least three integral U5 snRNP proteins thus potentially facilitate conformational changes in the spliceosome during nuclear pre-mRNA splicing.

NCBI



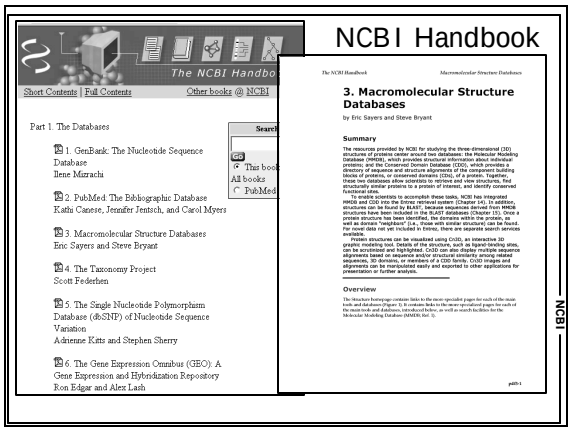
Books

NCBI



Online Books

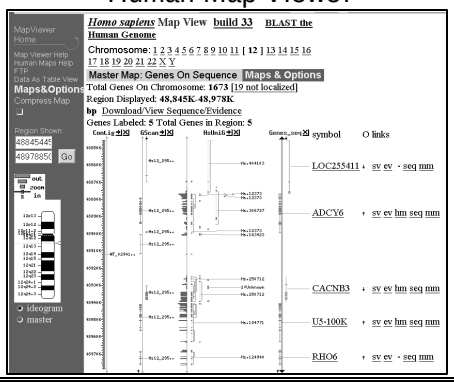
NCBI



NCBI Handbook

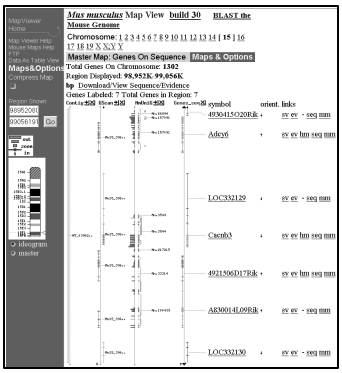
NCBI

Human Map Viewer



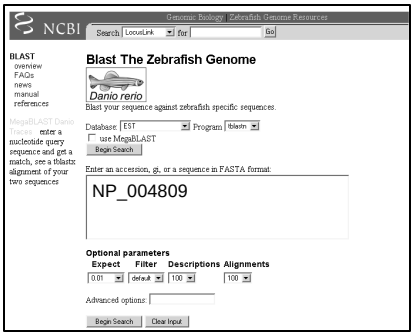
NCBI

Mouse

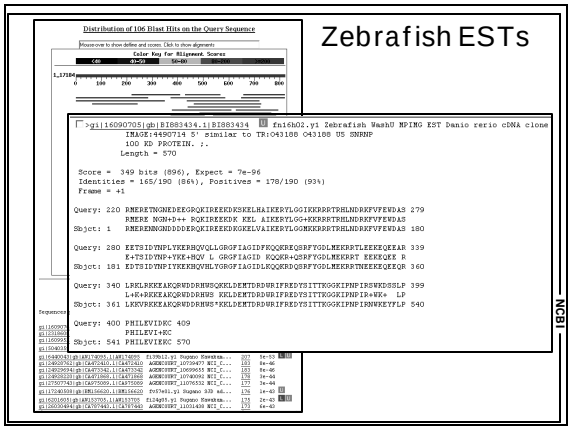


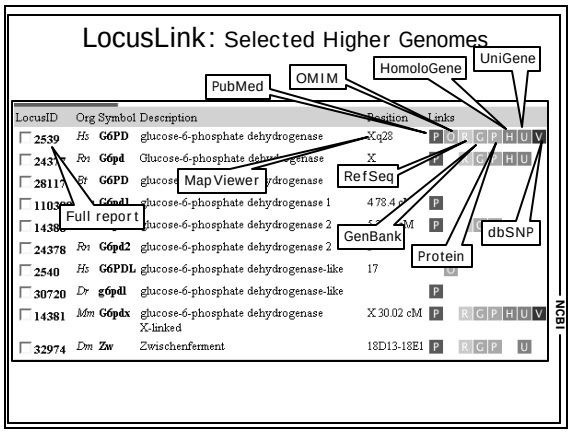
NCBI

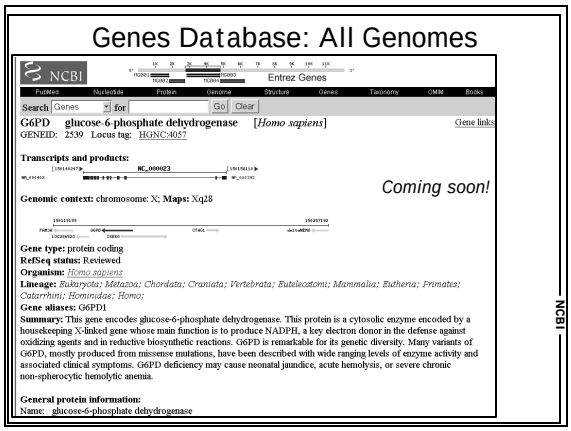
Zebrafish BLAST



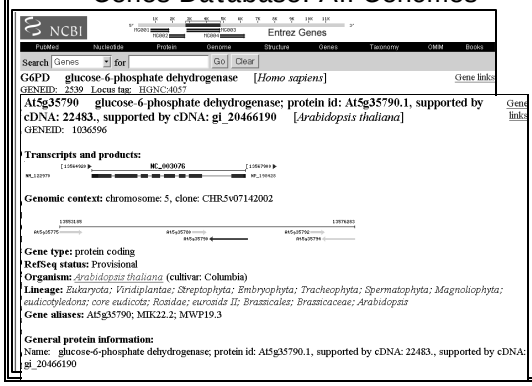
NCBI


Zebrafish ESTs
Distribution of 106 Blast Hits on the Query Sequence
NCBI


LocusLink: Selected Higher Genomes
PubMed
OMIM
HomoloGene
UniGene
MapViewer
RefSeq
Full report
GenBank
Protein
dbSNP
LocusID Org Symbol Description Position Links
2539 Hs G6PD glucose-6-phosphate dehydrogenase Xq28
24378 Rn G6pd2 glucose-6-phosphate dehydrogenase 2
2540 Hs G6PDL glucose-6-phosphate dehydrogenase-like 17
30720 Dr g6pdl glucose-6-phosphate dehydrogenase-like
14381 Mm G6pdx glucose-6-phosphate dehydrogenase X-linked X 30.02 cM
32974 Dm Zw Zwischenferment 18D13-18E1
NCBI

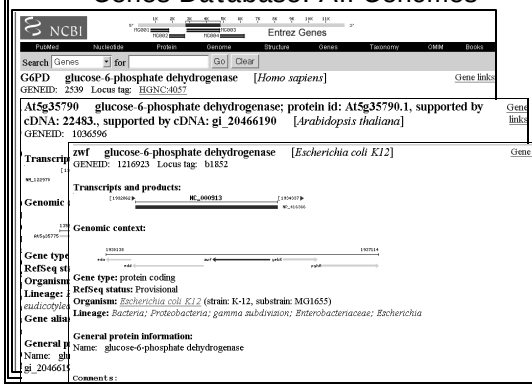

Genes Database: All Genomes
NCBI
Entrez Genes
PubMed Nucleotide Protein Genome Structure Genes Taxonomy OMIM Books
Search Genes for Go Clear
G6PD glucose-6-phosphate dehydrogenase [Homo sapiens]
Gene links
GENEID: 2539 Locus tag: HGNC:4057
Transcripts and products:
Coming soon!
Genomic context: chromosome: X; Maps: Xq28
Gene type: protein coding
RefSeq status: Reviewed
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo;
Gene aliases: G6PD1
Summary: This gene encodes glucose-6-phosphate dehydrogenase. This protein is a cytosolic enzyme encoded by a housekeeping X-linked gene whose main function is to produce NADPH, a key electron donor in the defense against oxidizing agents and in reductive biosynthetic reactions. G6PD is remarkable for its genetic diversity. Many variants of G6PD, mostly produced from missense mutations, have been described with wide ranging levels of enzyme activity and associated clinical symptoms. G6PD deficiency may cause neonatal jaundice, acute hemolysis, or severe chronic non-spherocytic hemolytic anemia.
General protein information:
Name: glucose-6-phosphate dehydrogenase
NCBI

Genes Database: All Genomes



NCBI

Genes Database: All Genomes



NCBI

Service Addresses

•***General Help***	`info@ncbi.nlm.nih.gov`
•***BLAST***	`blast-help@ncbi.nlm.nih.gov`

NCBI