NCBI Mini-Course

Unmasking Genes in Human DNA

Dr. Medha Bhagwat, NCBI
(bhagwat@ncbi.nlm.nih.gov)

Dr. David Wheeler, NCBI
(wheeler@ncbi.nlm.nih.gov)

NCBI Mini-Course
Unmasking Genes in Human DNA

(http://www.ncbi.nlm.nih.gov/Class/wheeler/Javagene/gg.html)

This course is an introduction to mining the human genome.

First, we will predict the exons in the protein-coding genes using two gene prediction tools:

1. GenScan (http://genes.mit.edu/GENSCAN.html)
2. GeneMark (http//dixie.biology.gatech.edu/GeneMark/eukhmm.cgi)

in conjunction with

1. BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/)
2. BLASTN against EST database. (http://www.ncbi.nlm.nih.gov/BLAST/)
3. Comparison with the mouse and rat genomic sequences (http://www.ncbi.nlm.nih.gov/BLAST/)

In addition, we will also predict the presence of

1. Promoters using PROSCAN (http://bimas.dcrt.nih.gov/molbio/proscan/)

2. Repeat elements using RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker)

We will then assemble the amino acid sequence of the gene product, on the basis of the exons chosen, using the translation tool provided on the class web page.

The web page contains links to all the necessary tools for the analysis. It also includes an ability to translate the DNA sequence into amino acid sequence by selecting the appropriate exon sequences. During the first hour, an instructor will walk you through an analysis of some genomic sequences. During the second hour of the class, you will perform the same analysis using different genomic sequences that will be provided to you.

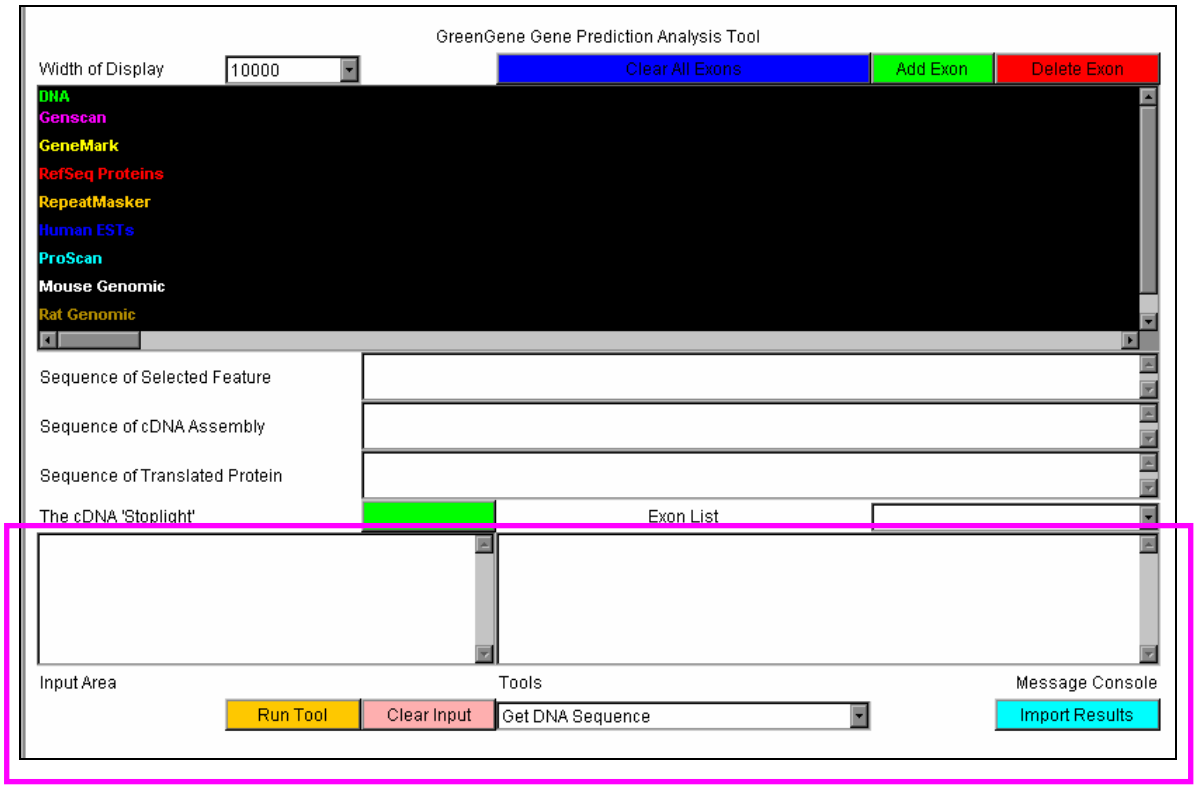
Greengene is a Java application which accesses several web-based sequence-analysis tools relevant to the identification of eukaryotic genes, then captures and integrates their output in a single view for ease of comprehension. Using **Greengene**, exons can be picked interactively and assembled into a coding sequence, then translated into a protein product. The exon choices made by the user reflect the information provided by several of the tools used rather than that of a single tool and are, therefore, more reliable. The tools accessed are GenScan and Genemark (exon prediction),

Repeatmasker (repeat identification), Proscan (promoter prediction), and blastx and blastn (to support exon prediction).

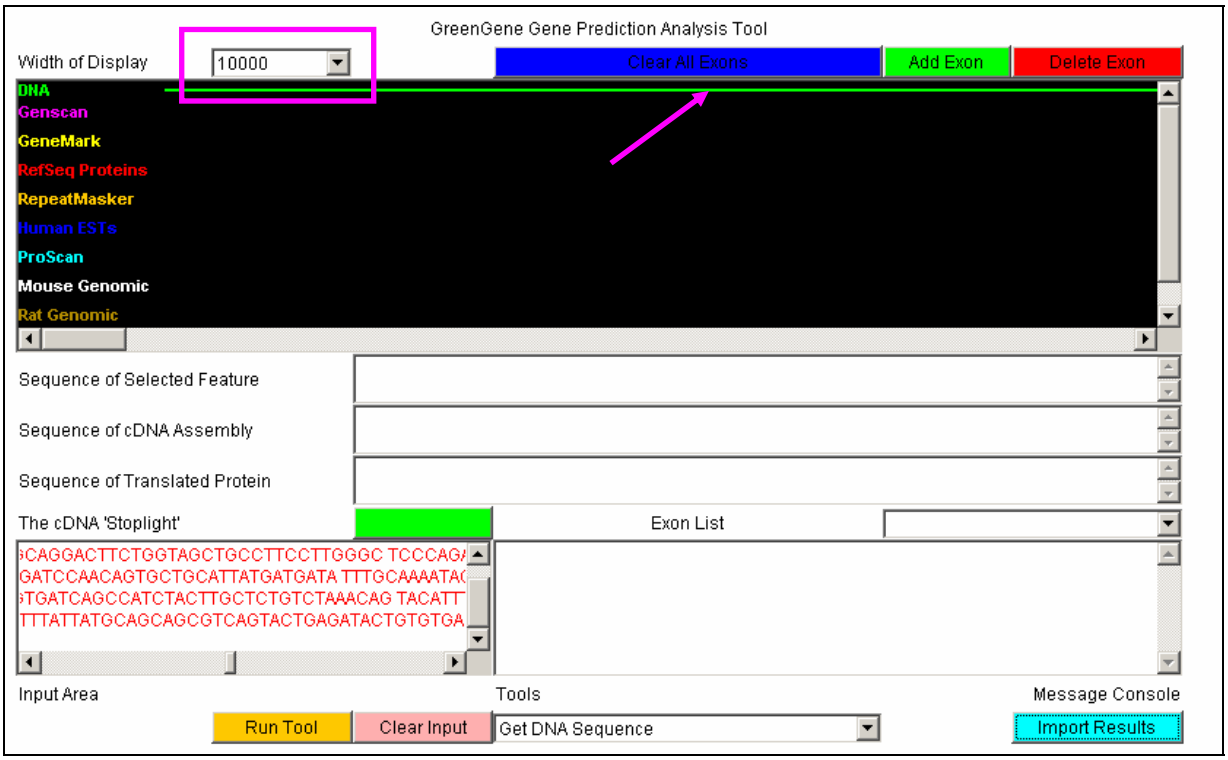
At least 50% of the human genome sequence is made up of repeat elements. Most of the eukaryotic protein coding genes contain exons and non-coding introns. The exons make up only about 1% of the human genome.

To correctly identify the exons in a eukaryotic gene you will need to compare the output of at least two different gene-prediction programs and couple this with the outputs given by a blastx search against protein sequences and a blastn search against ESTs. Greengene performs the data integration for you. The most reliable exons should be predicted by both programs and should align with blastx and/or blastn hits. In some cases, an exon prediction program may generate a potentially spurious exon without blastx or blastn support. In this case, you may want to exclude this exon in your gene model assembly. **Greengene** will allow you to do this easily so that you can create a custom coding sequence and the corresponding protein sequence without having to accept the automatically generated output of any single program.

Sequence-analysis tools are selected from a list-box labeled “Tools”. Once selected, a tool can either be “Run” or its output can be “Imported” into **Greengene**. When a tool is run, a second browser window is opened to the input page of the tool. The sequence to be analyzed must then be pasted into the tool’s input box and the tool must be run. When the tool has returned its output, the entire output should be selected (Select All), and pasted in the “Input Area”. To import the data, the “Import Results” button is then pressed.



To import one of the example sequences, select the “Sequence” tool, click on “Run”, highlight and copy the sequence desired, paste it in the “Input Area” and click on “Import Results.” You may paste DNA sequence of your choice not provided in the Sequence file.

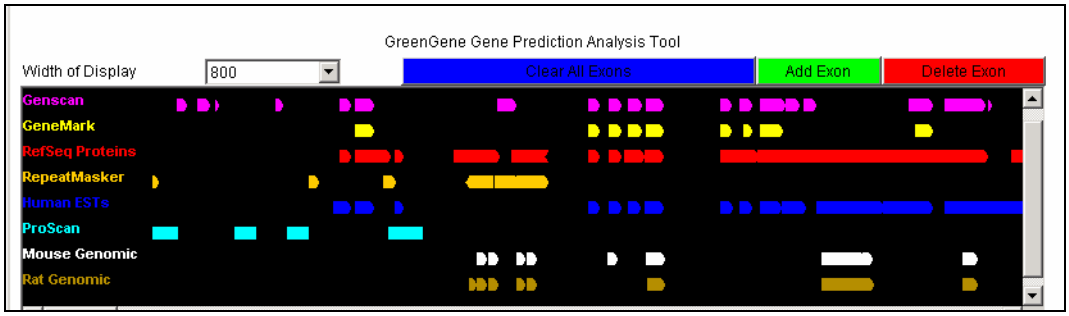


Once the import operation is complete, colored blocks will appear on the line corresponding to the proper tool in the display pane under the green line representing the DNA sequence. These blocks give the locations of features returned by the analysis tool. The width of the Display panel can be adjusted using the list box. Select 800 to zoom out to the full length of the DNA sequence.

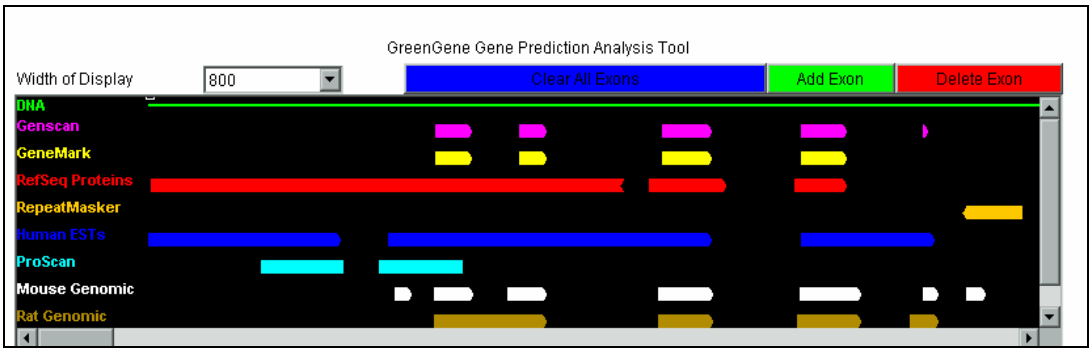
To get information about a feature, click on the colored block. A click prints information in the “Message Console” area and puts the sequence of the feature in question into the “Sequence of Selected Feature” area. To put the entire DNA sequence into the “Sequence of Selected Feature” area, as required for pasting into an analysis tool, just click the green sequence line.

To assemble a coding sequence, click on the first exon you want to include from the GenScan or the GeneMark output and press the “Add Exon” button. Continue doing this until you have added the sequences of all the exons you wish to include in your gene model. The cDNA sequence of the assembled gene model is shown in the “Sequence of cDNA Assembly” box and its amino acid translation in the first reading frame in the “Sequence of Translated Protein”. Look out for asterisks as they indicate stops and should be present only at the end of your protein translation. Internal stops indicate a problem in your exon assembly and are indicated by red color in “The cDNA stoplight” whereas a terminal stop codon is indicated by grey color. You can get to the list of all exons used in the assembly from the “Exon List” pull down menu. You may delete an exon by selecting it from the list and clicking on “Delete Exon. There is also a button to “Clear All Exons” and start over the assembly.

Below are **Greengene** outputs for 2 example DNA sequences that will be used in the class. To analyze one of these yourself, select “Initial Sequence” and hit ‘Run”. This will open a browser window containing the initial sequences to use as well as complete outputs from the various tools to use in case of web problems. We will use DNA2 during the demonstration; run the entire analysis on DNA1 yourself.



DNA1: Note the 7th feature in the Genscan track; is this likely to be a real exon?



DNA2: All exon predictions supported by blastx and blastn hits.

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

```
              \\|//
              (o o)
-. .-.   .-oOOo-(_)-oOOo-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.
||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\
|/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||
'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-'
```

For information about Genscan, click here

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

News:

- **The server is back online. Apologies for the delay.**
  - The interface now supports file upload, allowing you to input long sequences without cutting and pasting. The old cut-and-paste method also works.
  - Also, the server now outputs PDF (portable document format) images of predicted gene locations, as well as Postscript.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored): Browse

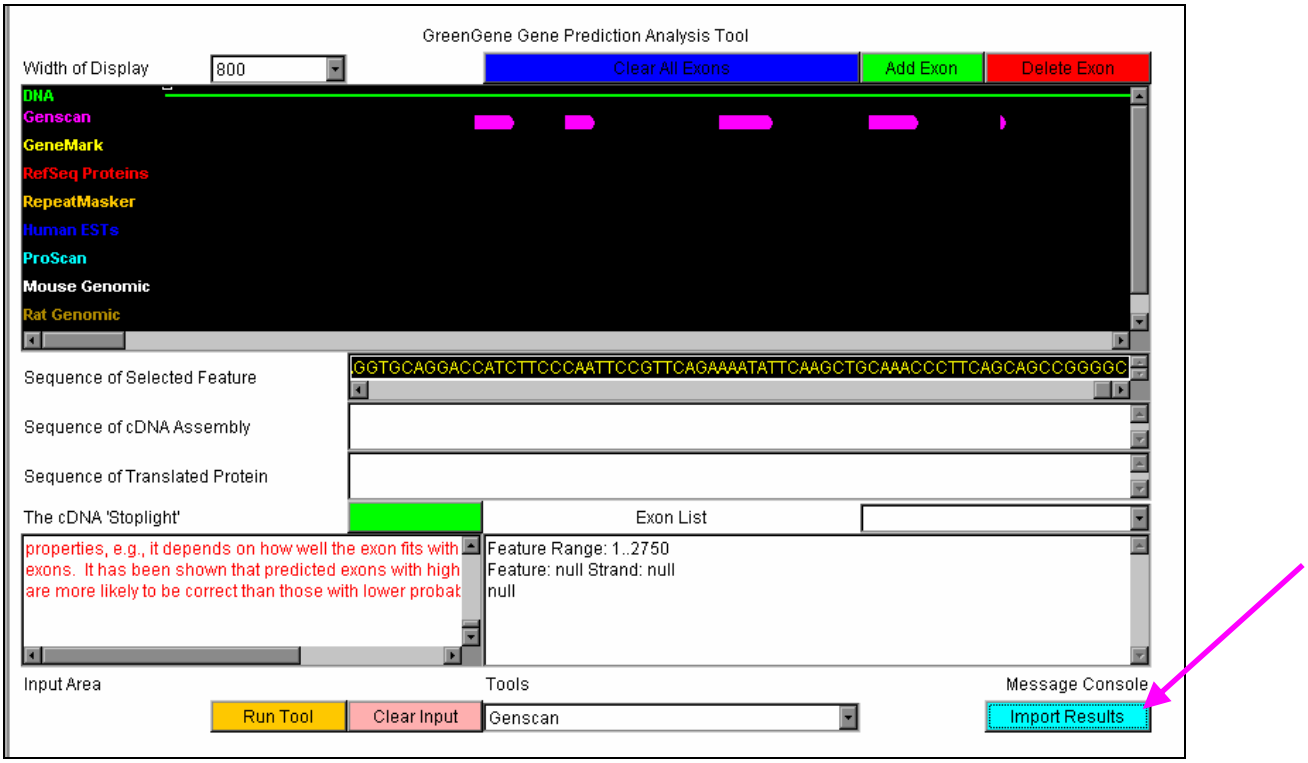
Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

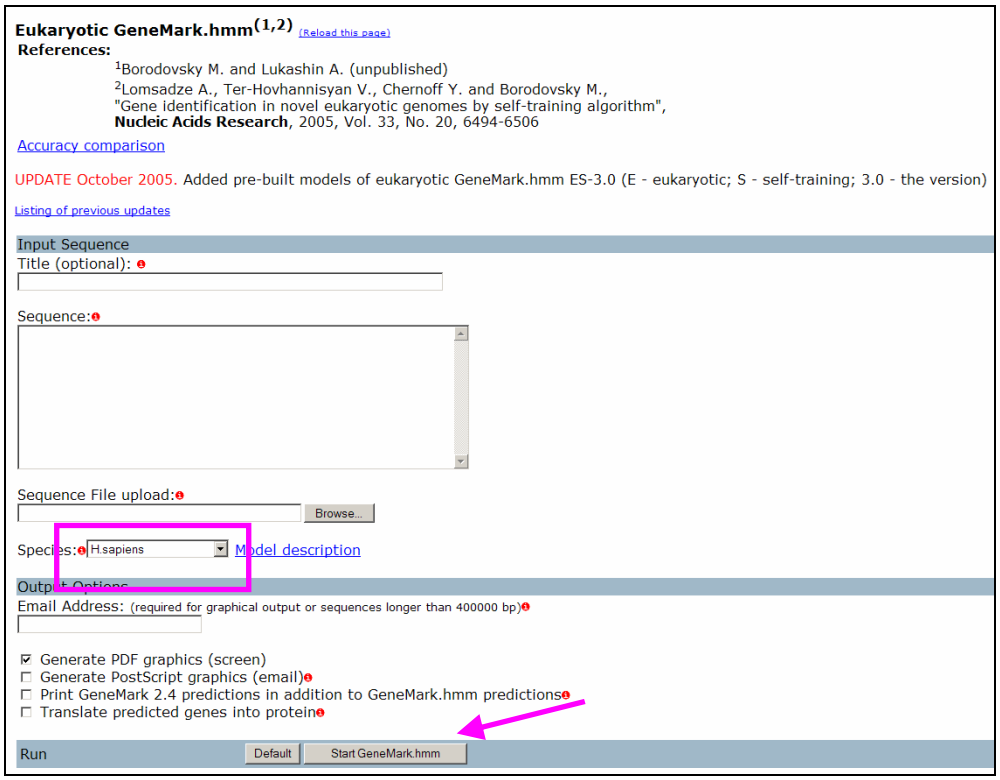
```
TCCCAAAGGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC
```

To have the results mailed to you, enter your email address here (optional):

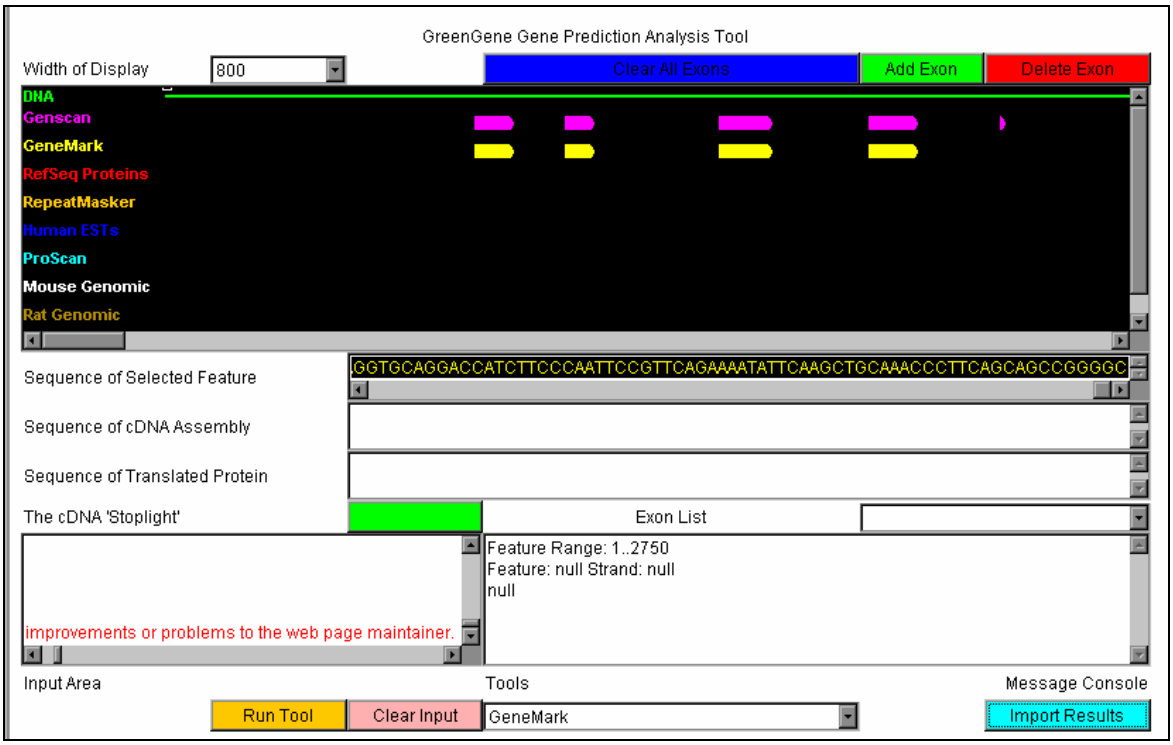
Run GENSCAN Clear Input

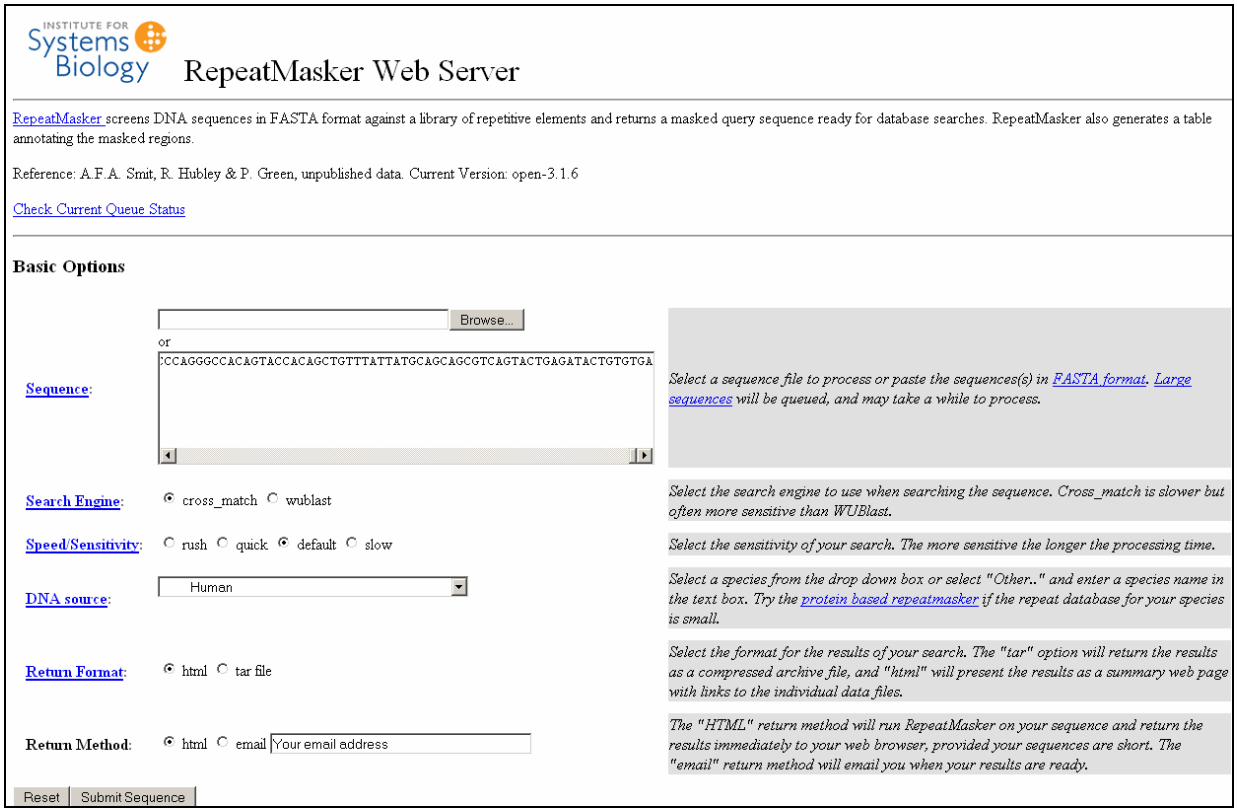
```
GENSCAN 1.0     Date run: 22-Sep-104    Time: 13:26:04

Sequence 13:26:04 : 2750 bp : 57.42% C+G : Isochore 4 (57 - 100 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:


Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Intr +    862    960   99  0  0   84   51   168 0.956  14.17
 1.02 Intr +   1113   1184   72  2  0   91   76   108 0.978  10.65
 1.03 Intr +   1542   1679  138  2  0  107   61   184 0.998  19.44
 1.04 Term +   1958   2086  129  1  0   78   48    92 0.906   3.32
 1.05 PlyA +   2323   2328    6                                1.05
```






```
GeneMark.hmm (Version 2.2a)
Sequence name: Fri Feb  9 11:30:34 EST 2007
Sequence length: 2750 bp
G+C content: 57.42%
Matrix: Homo sapiens
Fri Feb  9 11:30:34 2007

Predicted genes/exons

Gene Exon Strand Exon          Exon Range     Exon      Start/End
  #    #         Type                         Length       Frame

  1     1   +  Internal      862       960      99          1 3
  1     2   +  Internal     1113      1184      72          1 3
  1     3   +  Internal     1542      1679     138          1 3
  1     4   +  Terminal     1958      2086     129          1 3
```


GreenGene Gene Prediction Analysis Tool
Width of Display
800
Clear All Exons
Add Exon
Delete Exon
DNA
Genscan
GeneMark
RefSeq Proteins
RepeatMasker
Human ESTs
ProScan
Mouse Genomic
Rat Genomic
Sequence of Selected Feature
GGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC
Sequence of cDNA Assembly
Sequence of Translated Protein
The cDNA 'Stoplight'
Exon List
Feature Range: 1..2750
Feature: null Strand: null
null
improvements or problems to the web page maintainer.
Input Area
Tools
Message Console
Run Tool
Clear Input
GeneMark
Import Results

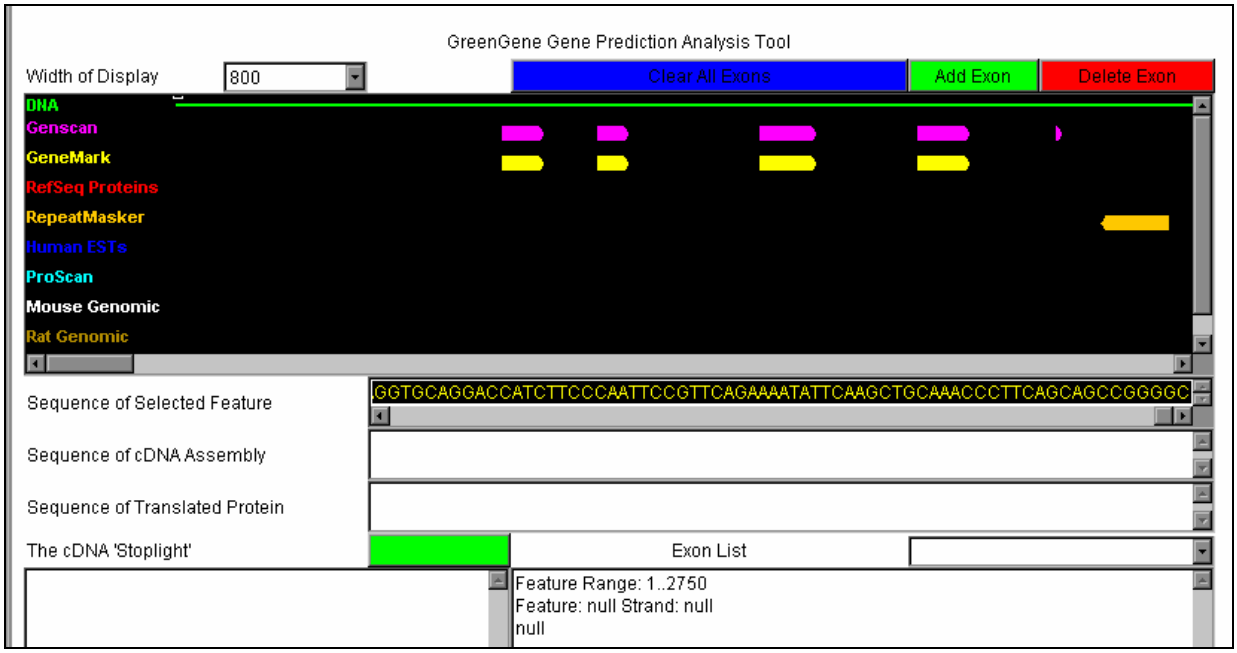

INSTITUTE FOR Systems Biology
RepeatMasker Web Server
RepeatMasker screens DNA sequences in FASTA format against a library of repetitive elements and returns a masked query sequence ready for database searches. RepeatMasker also generates a table annotating the masked regions.
Reference: A.F.A. Smit, R. Hubley & P. Green, unpublished data. Current Version: open-3.1.6
Check Current Queue Status
Basic Options
Browse...
or
CCAGGGCCACAGTACCACAGCTGTTTATTATGCAGCAGCGTCAGTACTGAGATACTGTGTGA
Sequence:
Select a sequence file to process or paste the sequences(s) in FASTA format. Large sequences will be queued, and may take a while to process.
Search Engine:
cross_match
wublast
Select the search engine to use when searching the sequence. Cross_match is slower but often more sensitive than WUBlast.
Speed/Sensitivity:
rush
quick
default
slow
Select the sensitivity of your search. The more sensitive the longer the processing time.
DNA source:
Human
Select a species from the drop down box or select "Other.." and enter a species name in the text box. Try the protein based repeatmasker if the repeat database for your species is small.
Return Format:
html
tar file
Select the format for the results of your search. The "tar" option will return the results as a compressed archive file, and "html" will present the results as a summary web page with links to the individual data files.
Return Method:
html
email
Your email address
The "HTML" return method will run RepeatMasker on your sequence and return the results immediately to your web browser, provided your sequences are short. The "email" return method will email you when your results are ready.
Reset
Submit Sequence

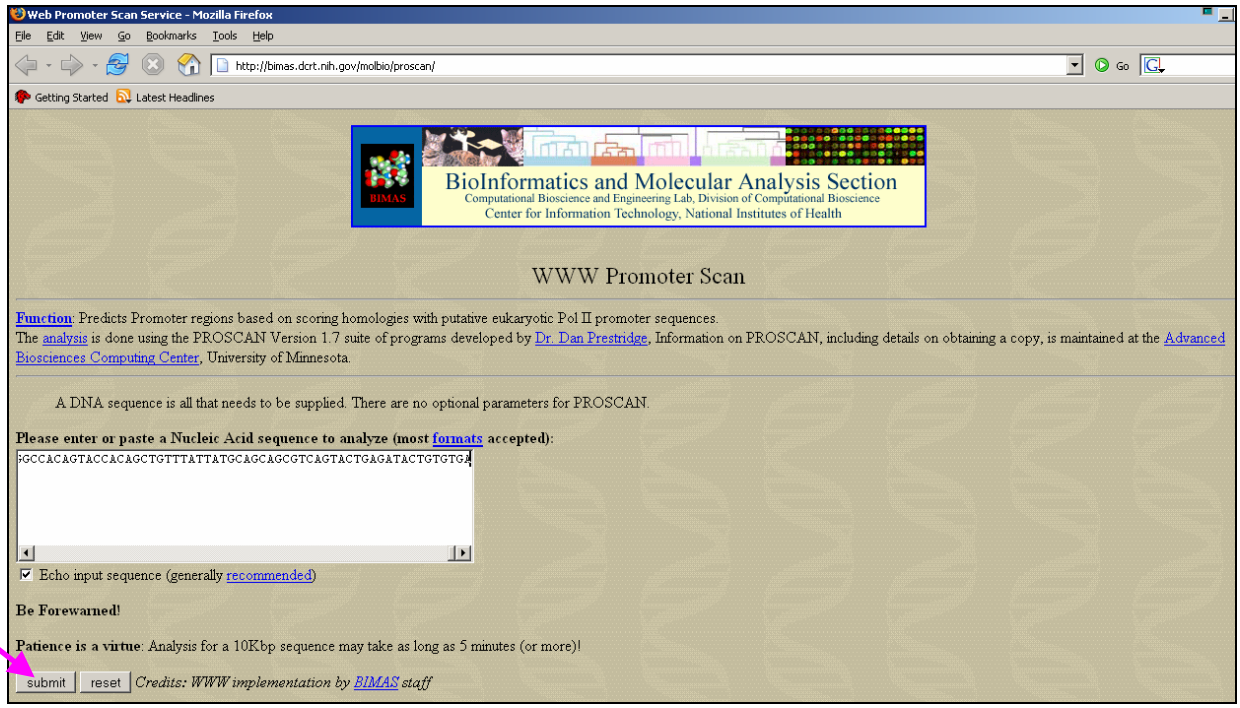
Results

Right-click and select "Save As" to save results to your computer or click on the link to view the file in the browser.

Annotation File: RM2sequpload_1187105682.out.html (NEW XHTML Format)
RM2sequpload_1187105682.out.txt (Text File Format)
Masked File: RM2sequpload_1187105682.masked
Alignment File: RM2sequpload_1187105682.align

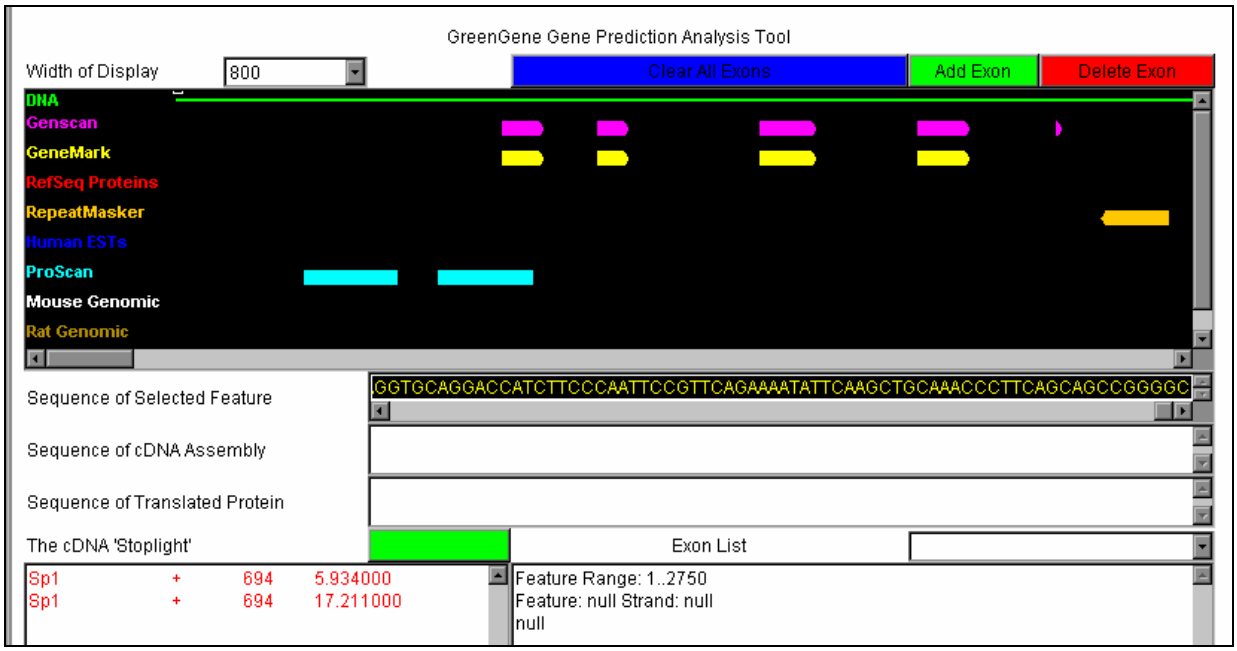
SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin	end	(left)		matching repeat repeat	class/family	position in repeat begin	end	(left)	ID
607	17.7	11.2	0.5	UnnamedSequence	2453	2622	(128)	C	MER20	DNA/MER1_type	(2)	217	30	1

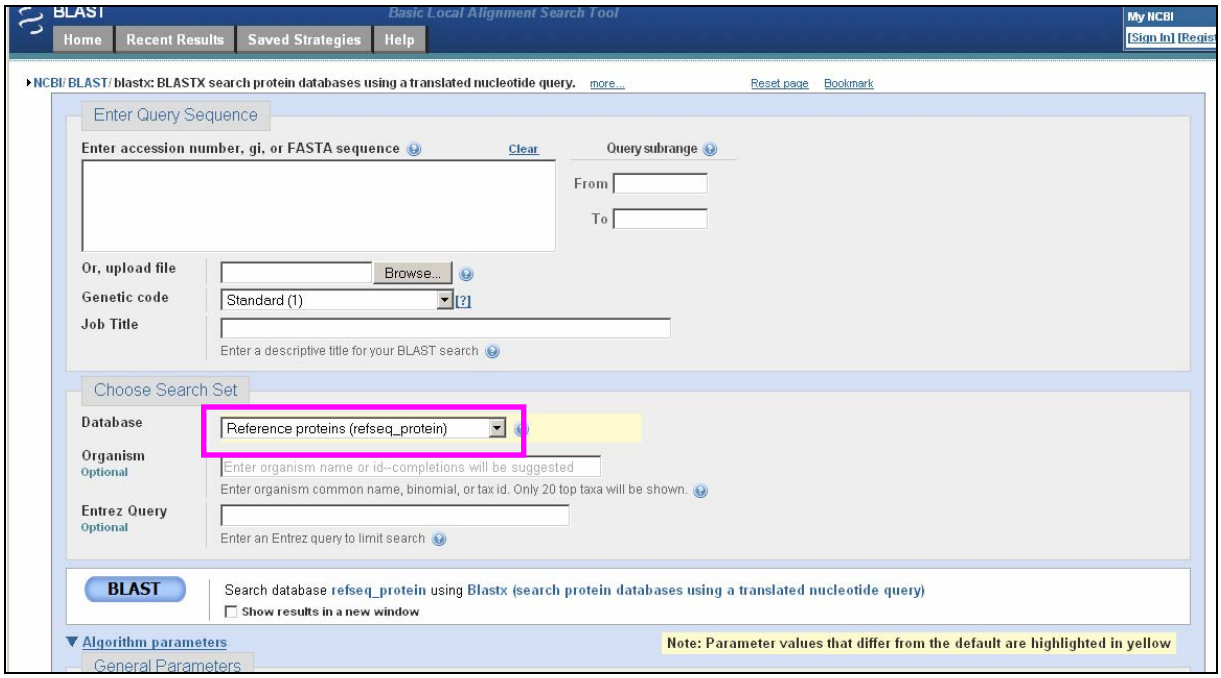




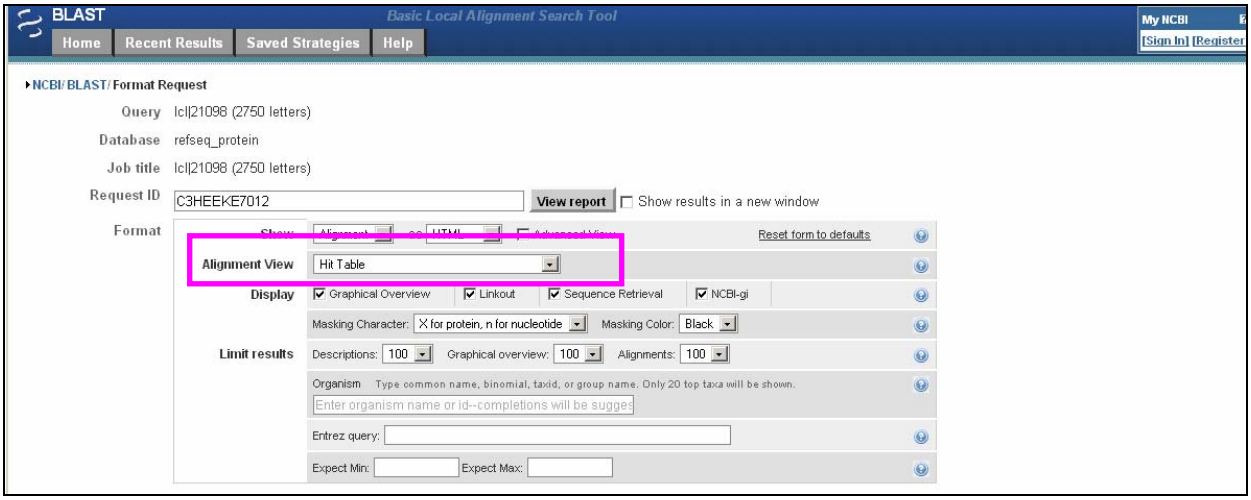
Significant Signals:

Name	TFD #	Strand	Location	Weight
HSV_IE_repeat	S01565	-	351	1.363000
Sp1	S01542	-	354	3.608000
JCV_repeated_sequenc	S01193	-	364	1.658000
T-Ag	S00974	+	398	1.086000
NF-kB	S01498	-	432	1.008000
EARLY-SEQ1	S01081	+	472	6.322000
Sp1	S00801	+	473	2.755000
Sp1	S00802	+	474	3.292000
Sp1	S00781	-	478	2.772000
Sp1	S00978	-	479	3.361000
JCV_repeated_sequenc	S01193	-	479	1.658000
Sp1	S00952	+	490	50.000000
Sp1	S01542	-	499	3.608000
beta-pol_CS	S00559	+	510	8.603000
ATF	S01059	+	511	1.157000
CREB	S00969	+	511	3.442000
CREB	S00072	+	511	8.603000


GreenGene Gene Prediction Analysis Tool
Width of Display
800
Clear All Exons
Add Exon
Delete Exon
DNA
Genscan
GeneMark
RefSeq Proteins
RepeatMasker
Human ESTs
ProScan
Mouse Genomic
Rat Genomic
Sequence of Selected Feature
GGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC
Sequence of cDNA Assembly
Sequence of Translated Protein
The cDNA 'Stoplight'
Exon List
Sp1 + 694 5.934000
Sp1 + 694 17.211000
Feature Range: 1..2750
Feature: null Strand: null
null


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Regist
NCBI/ BLAST/ blastx: BLASTX search protein databases using a translated nucleotide query.
more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Query subrange
From
To
Or, upload file
Browse...
Genetic code
Standard (1)
[?]
Job Title
Enter a descriptive title for your BLAST search
Choose Search Set
Database
Reference proteins (refseq_protein)
Organism
Optional
Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Entrez Query
Optional
Enter an Entrez query to limit search
BLAST
Search database refseq_protein using Blastx (search protein databases using a translated nucleotide query)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow
General Parameters

BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Regis
NCBI/ BLAST/ blastx/ Formatting Results - C3HEEKE7012
[Formatting options]
Job Title: lcl|21098 (2750 letters)
Request ID
C3HEEKE7012
Status
Searching
Submitted at
Tue Aug 14 11:10:06 2007
Current time
Tue Aug 14 11:10:31 2007
Time since submission
This page will be automatically updated in 3 seconds


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register
NCBI/ BLAST/ Format Request
Query
lcl|21098 (2750 letters)
Database
refseq_protein
Job title
lcl|21098 (2750 letters)
Request ID
C3HEEKE7012
View report
Show results in a new window
Format
Reset form to defaults
Alignment View
Hit Table
Display
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Masking Character:
X for protein, n for nucleotide
Masking Color:
Black
Limit results
Descriptions:
100
Graphical overview:
100
Alignments:
100
Organism
Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.
Enter organism name or id--completions will be sugges
Entrez query:
Expect Min:
Expect Max:

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/BLAST/blastx/Formatting Results - C3HEEKE7012 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search strategy]

Job Title: lcl|21098 (2750 letters)

```
# BLASTX 2.2.17 (Jun-24-2007)
# Query:
# Database: refseq_protein
# Fields: query id, subject ids, % identity, % positives, query/sbjct frames, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue,
# 228 hits found
21098   gi|109113167|ref|XP_001111216.1|        82.86   87.14   3/0     70      12      0       984     1193    42      111     2e-24   118
21098   gi|109113167|ref|XP_001111216.1|        67.12   71.23   1/0     73      22      1       862     1080    1       71      3e-19   100
21098   gi|109113167|ref|XP_001111216.1|        97.87   100.00  3/0     47      1       0       1542    1682    109     155     6e-19   99.8
21098   gi|109113167|ref|XP_001111216.1|        100.00  100.00  2/0     42      0       0       1958    2083    155     196     7e-15   86.3
21098   gi|73955588|ref|XP_859762.1|    90.38   94.23   3/0     52      5       0       1527    1682    18      69      4e-19   100
21098   gi|73955588|ref|XP_859762.1|    100.00  100.00  2/0     42      0       0       1958    2083    69      110     7e-15   86.3
21098   gi|73955588|ref|XP_859762.1|    72.73   86.36   2/0     22      6       0       896     961     1       22      6.6     36.6
21098   gi|10864021|ref|NP_067033.1|;gi|115497316|ref|NP_001069881.1|;gi|73955584|ref|XP_536628.2|;gi|73955590|ref|XP_859798.1|;gi|109113163|ref|XP_001111285.1|;gi
21098   gi|10864021|ref|NP_067033.1|;gi|115497316|ref|NP_001069881.1|;gi|73955584|ref|XP_536628.2|;gi|73955590|ref|XP_859798.1|;gi|109113163|ref|XP_001111285.1|;gi
21098   gi|10864021|ref|NP_067033.1|;gi|115497316|ref|NP_001069881.1|;gi|73955584|ref|XP_536628.2|;gi|73955590|ref|XP_859798.1|;gi|109113163|ref|XP_001111285.1|;gi
21098   gi|10864021|ref|NP_067033.1|;gi|115497316|ref|NP_001069881.1|;gi|73955584|ref|XP_536628.2|;gi|73955590|ref|XP_859798.1|;gi|109113163|ref|XP_001111285.1|;gi
21098   gi|66730565|ref|NP_001019377.1|;gi|86439955|ref|NP_001034467.1| 95.74   100.00  3/0     47      2       0       1542    1682    58      104     1e-18   98.
21098   gi|66730565|ref|NP_001019377.1|;gi|86439955|ref|NP_001034467.1| 97.62   97.62   2/0     42      1       0       1958    2083    104     145     2e-14   84.
21098   gi|66730565|ref|NP_001019377.1|;gi|86439955|ref|NP_001034467.1| 100.00  100.00  1/0     33      0       0       862     960     1       33      4e-12   77.
21098   gi|66730565|ref|NP_001019377.1|;gi|86439955|ref|NP_001034467.1| 89.29   92.86   3/0     28      3       0       1110    1193    33      60      9e-05   52.
21098   gi|126309130|ref|XP_001364568.1|        93.62   97.87   3/0     47      3       0       1542    1682    58      104     2e-18   97.8
21098   gi|126309130|ref|XP_001364568.1|        97.62   100.00  2/0     42      1       0       1958    2083    104     145     2e-14   85.1
21098   gi|126309130|ref|XP_001364568.1|        96.97   100.00  1/0     33      1       0       862     960     1       33      2e-11   75.1
21098   gi|126309130|ref|XP_001364568.1|        85.71   92.86   3/0     28      4       0       1110    1193    33      60      2e-04   51.6
21098   gi|149438415|ref|XP_001508283.1|        87.23   95.74   3/0     47      6       0       1542    1682    58      104     1e-16   92.0
21098   gi|149438415|ref|XP_001508283.1|        96.97   96.97   1/0     33      1       0       862     960     1       33      2e-11   75.1
21098   gi|149438415|ref|XP_001508283.1|        82.93   92.68   2/0     41      7       0       1958    2080    104     144     2e-11   74.7
21098   gi|149438415|ref|XP_001508283.1|        89.29   92.86   3/0     28      3       0       1110    1193    33      60      5e-05   53.5
21098   gi|73955586|ref|XP_859732.1|    100.00  100.00  2/0     42      0       0       1958    2083    94      135     7e-15   86.3
21098   gi|73955586|ref|XP_859732.1|    97.30   100.00  3/0     37      1       0       1572    1682    58      94      5e-13   80.1
21098   gi|73955586|ref|XP_859732.1|    100.00  100.00  1/0     33      0       0       862     960     1       33      4e-12   77.0
21098   gi|73955586|ref|XP_859732.1|    92.00   100.00  3/0     25      2       0       1110    1184    33      57      4e-04   50.4
21098   gi|148235479|ref|NP_001086138.1|        72.34   89.36   3/0     47      13      0       1542    1682    58      104     3e-13   80.9
21098   gi|148235479|ref|NP_001086138.1|        78.79   96.97   1/0     33      7       0       862     960     1       33      5e-09   67.0
21098   gi|148235479|ref|NP_001086138.1|        60.42   81.25   2/0     48      19      0       1940    2083    98      145     2e-08   65.1
21098   gi|148235479|ref|NP_001086138.1|        82.14   89.29   3/0     28      5       0       1110    1193    33      60      0.002   48.1
21098   gi|56118869|ref|NP_001008076.1| 72.34   89.36   3/0     47      13      0       1542    1682    58      104     3e-13   80.9
21098   gi|56118869|ref|NP_001008076.1| 78.79   96.97   1/0     33      7       0       862     960     1       33      3e-09   67.4
21098   gi|56118869|ref|NP_001008076.1| 60.42   81.25   2/0     48      19      0       1940    2083    98      145     2e-08   65.1
21098   gi|56118869|ref|NP_001008076.1| 78.57   85.71   3/0     28      6       0       1110    1193    33      60      0.006   46.6
21098   gi|147905298|ref|NP_001088916.1|        70.21   87.23   3/0     47      14      0       1542    1682    58      104     5e-13   80.1
21098   gi|147905298|ref|NP_001088916.1|        78.79   96.97   1/0     33      7       0       862     960     1       33      3e-09   67.4
21098   gi|147905298|ref|NP_001088916.1|        60.42   81.25   2/0     48      19      0       1940    2083    98      145     2e-08   65.1
21098   gi|147905298|ref|NP_001088916.1|        82.14   89.29   3/0     28      5       0       1110    1193    33      60      0.002   48.1
21098   gi|57526527|ref|NP_001002742.1| 68.09   80.85   3/0     47      15      0       1542    1682    58      104     1e-10   72.0
21098   gi|57526527|ref|NP_001002742.1| 87.88   93.94   1/0     33      4       0       862     960     1       33      1e-09   68.9
21098   gi|57526527|ref|NP_001002742.1| 70.73   92.68   2/0     41      12      0       1958    2080    104     144     3e-09   67.4
21098   gi|57526527|ref|NP_001002742.1| 85.71   92.86   3/0     28      4       0       1110    1193    33      60      3e-04   51.2
21098   gi|72071136|ref|XP_795111.1|;gi|115925633|ref|XP_001195648.1|   38.20   57.30   3/0     89      49      1       1542    1808    58      140     2e-09   68.
21098   gi|72071136|ref|XP_795111.1|;gi|115925633|ref|XP_001195648.1|   50.00   80.95   2/0     42      21      0       1958    2083    104     145     5e-05   53.
21098   gi|72071136|ref|XP_795111.1|;gi|115925633|ref|XP_001195648.1|   60.61   81.82   1/0     33      13      0       862     960     1       33      3e-04   51.
21098   gi|72071136|ref|XP_795111.1|;gi|115925633|ref|XP_001195648.1|   60.71   89.29   3/0     28      11      0       1110    1193    33      60      0.041   43.
21098   gi|125882528|ref|XP_001346257.1|        70.73   92.68   2/0     41      12      0       1958    2080    1       41      3e-09   67.4
21098   gi|154311447|ref|XP_001555053.1|        46.15   56.92   3/0     65      30      1       1545    1724    94      158     7e-07   59.7
```

GreenGene Gene Prediction Analysis Tool

Width of Display 800 | Clear All Exons | Add Exon | Delete Exon

DNA
Genscan
GeneMark
RefSeq Proteins
RepeatMasker
Human ESTs
ProScan
Mouse Genomic
Rat Genomic

Sequence of Selected Feature: GGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC

Sequence of cDNA Assembly

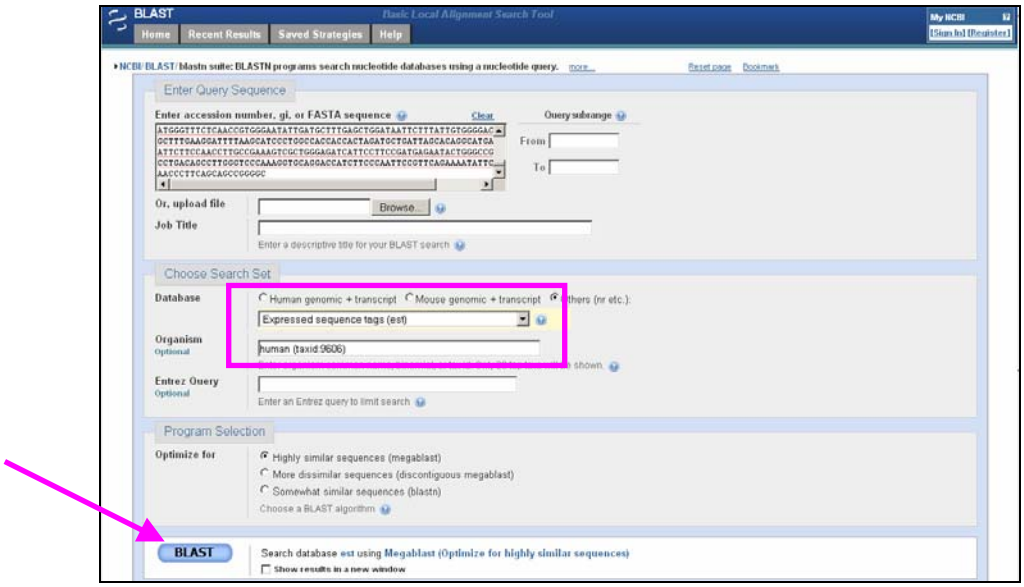
Sequence of Translated Protein

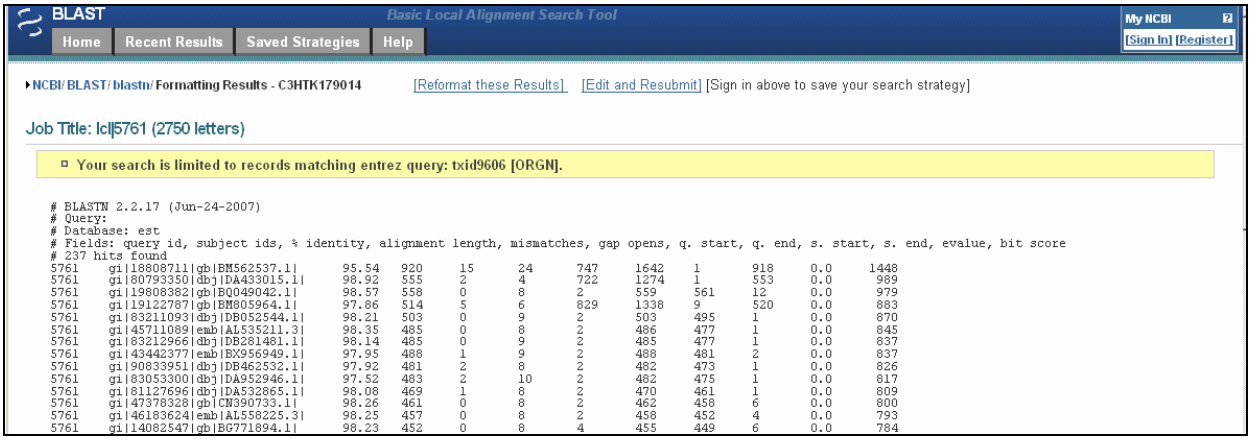
The cDNA 'Stoplight'

Exon List

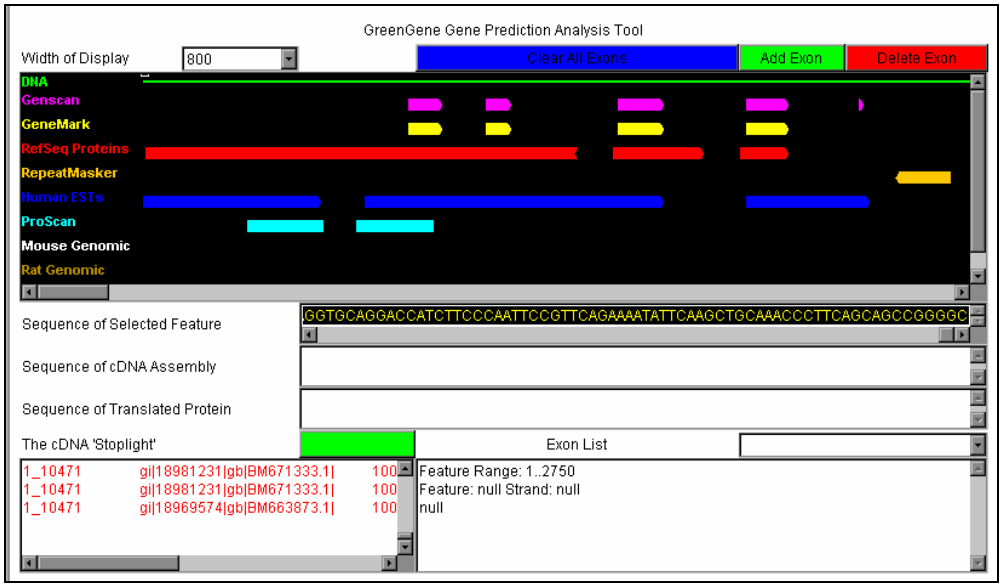
1_21789 gi|121537830|ref|ZP_01669606.1| 27.
1_21789 gi|121537830|ref|ZP_01669606.1| 25.
1_21789 gi|39940408|ref|XP_359741.1| 29.

Feature Range: 1..2750
Feature: null Strand: null
null






```
# BLASTN 2.2.17 (Jun-24-2007)
# Query:
# Database: est
# Fields: query id, subject ids, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 237 hits found
5761    gi|18808711|gb|BM562537.1|      95.54   920     15      24      747     1642    1       918     0.0     1448
5761    gi|80793350|dbj|DA433015.1|     98.92   555     2       4       722     1274    1       553     0.0     989
5761    gi|19808382|gb|BQ049042.1|      98.57   558     0       8       2       559     561     12      0.0     979
5761    gi|19122787|gb|BM805964.1|      97.86   514     5       6       829     1338    9       520     0.0     883
5761    gi|83211093|dbj|DB052544.1|     98.21   503     0       9       2       503     495     1       0.0     870
5761    gi|45711089|emb|AL535211.3|     98.35   485     0       8       2       486     477     1       0.0     845
5761    gi|83212966|dbj|DB281481.1|     98.14   485     0       9       2       485     477     1       0.0     837
5761    gi|43442377|emb|BX956949.1|     97.95   488     1       9       2       488     481     2       0.0     837
5761    gi|90833951|dbj|DB462532.1|     97.92   481     2       8       2       482     473     1       0.0     826
5761    gi|83053300|dbj|DA952946.1|     97.52   483     2       10      2       482     475     1       0.0     817
5761    gi|81127696|dbj|DA532865.1|     98.08   469     1       8       2       470     461     1       0.0     809
5761    gi|47378328|gb|CN390733.1|      98.26   461     0       8       2       462     458     6       0.0     800
5761    gi|46183624|emb|AL558225.3|     98.25   457     0       8       2       458     452     4       0.0     793
5761    gi|14082547|gb|BG771894.1|      98.23   452     0       8       4       455     449     6       0.0     784
```



NCBI Home ▸ Genomic Biology ▸ Mouse Genome Resources ▸ BLAST

Search Map Viewer Go Clear

BLAST
Overview
FAQs
News
Manual
References
Retrieve results

Genome Project

BLAST Mouse Sequences.

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

```
CCTGCCCCACGCATGAAGGTGGAAAGTAGGGCAGCCACACCCCCTGGAGGGCCACGAAGTGGCTT
ATGGGTTTCTCAACCGTGGGAATATTGATGCTTTGAGCTGGATAATTCTTTATTGTGGGGACCGT
GCTTTGAAGGATTTTAAGCATCCCTGGCCACCACCACTAGATGCTGATTAGCACAGGCATGAAAT
ATTCTTCCAACCTTGCCGAAAGTCGCTGGGAGATCATTCCTTCCGATGAGAATACTGGGCCGGGG
CCTGACAGCCTTGGGTCCCAAAGGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAG
AACCCTTCAGCAGCCGGGGC
```

Or, choose a file to upload

Browse...

Set subsequence: (optional)
From: To:

Database:
genome (reference only) 316 sequences

Program:
cross-species megaBLAST: Compare nucleotide sequences for other species to this genome

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options: -G 5 -E 2

Begin Search Clear Input

Get the URL with preset values ? Get URL

BLAST Basic Local Alignment Search Tool

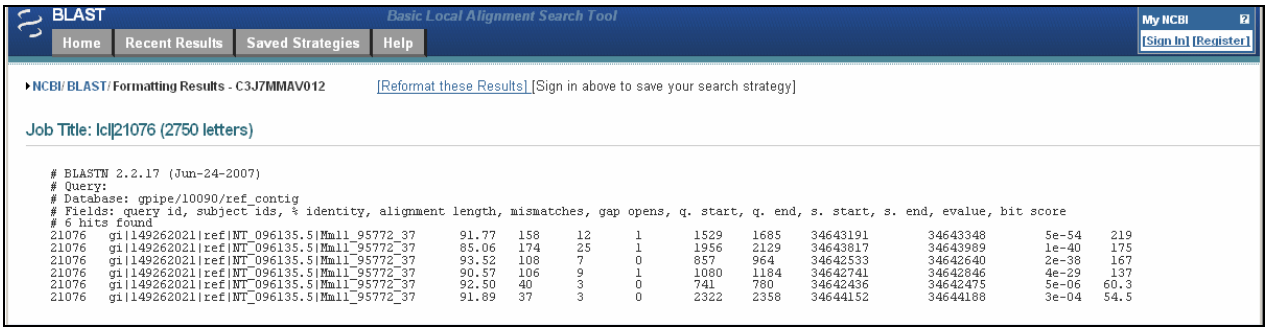
Home Recent Results Saved Strategies Help

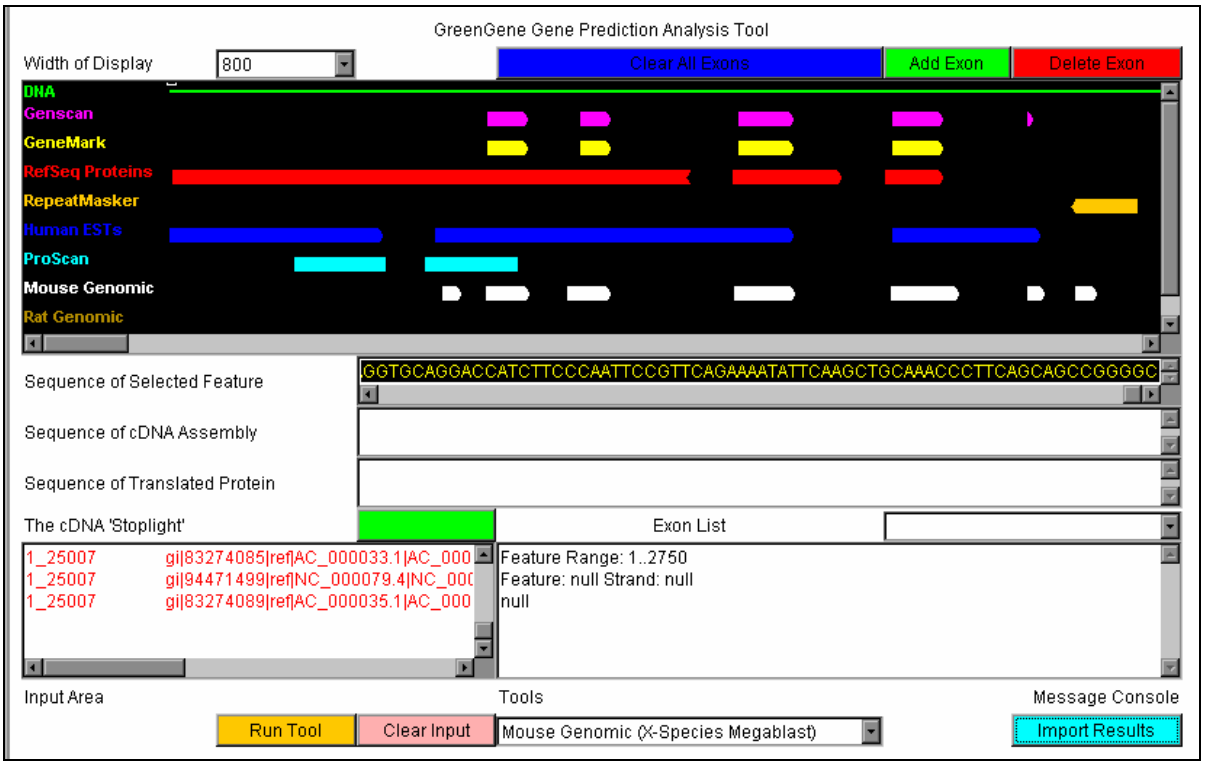
My NCBI [Sign In] [Register

NCBI/BLAST/Format Request

Query lcl|21076 (2750 letters)
Database ref_contig
Job title lcl|21076 (2750 letters)
Request ID C3J7MMAV012 View report Show results in a new window

Format
Show Alignment as HTML Advanced View Reset form to defaults
Alignment View Hit Table
Display Graphical Overview Linkout Sequence Retrieval NCBI-gi CDS feature
Masking Character: Lower Case Masking Color: Grey
Limit results Descriptions: 100 Graphical overview: 100 Alignments: 100
Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.
Enter organism name or id--completions will be sugges
Entrez query:
Expect Min: Expect Max:


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ Formatting Results - C3J7MMAV012
[Reformat these Results] [Sign in above to save your search strategy]
Job Title: lcl|21076 (2750 letters)
BLASTN 2.2.17 (Jun-24-2007)
Query:
Database: gpipe/10090/ref_contig
Fields: query id, subject ids, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
6 hits found
21076 gi|149262021|ref|NT_096135.5|Mm11_95772_37 91.77 158 12 1 1529 1685 34643191 34643348 5e-54 219
21076 gi|149262021|ref|NT_096135.5|Mm11_95772_37 85.06 174 25 1 1956 2129 34643817 34643989 1e-40 175
21076 gi|149262021|ref|NT_096135.5|Mm11_95772_37 93.52 108 7 0 857 964 34642533 34642640 2e-38 167
21076 gi|149262021|ref|NT_096135.5|Mm11_95772_37 90.57 106 9 1 1080 1184 34642741 34642846 4e-29 137
21076 gi|149262021|ref|NT_096135.5|Mm11_95772_37 92.50 40 3 0 741 780 34642436 34642475 5e-06 60.3
21076 gi|149262021|ref|NT_096135.5|Mm11_95772_37 91.89 37 3 0 2322 2358 34644152 34644188 3e-04 54.5


GreenGene Gene Prediction Analysis Tool
Width of Display
800
Clear All Exons
Add Exon
Delete Exon
DNA
Genscan
GeneMark
RefSeq Proteins
RepeatMasker
Human ESTs
ProScan
Mouse Genomic
Rat Genomic
Sequence of Selected Feature
GGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC
Sequence of cDNA Assembly
Sequence of Translated Protein
The cDNA 'Stoplight'
Exon List
1_25007 gi|83274085|ref|AC_000033.1|AC_000
1_25007 gi|94471499|ref|NC_000079.4|NC_00(
1_25007 gi|83274089|ref|AC_000035.1|AC_000
Feature Range: 1..2750
Feature: null Strand: null
null
Input Area
Tools
Message Console
Run Tool
Clear Input
Mouse Genomic (X-Species Megablast)
Import Results

NCBI Home ▸ Genomic Biology ▸ Rat Genome Resources ▸ BLAST

Search Map Viewer Go Clear

BLAST
Overview
FAQs
News
Manual
References
Retrieve results

Genome Project

BLAST Rat Sequences.

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

```
CCTGCCCCACGCATGAAGGTGGAAAGTAGGGCAGCCACACCCCCTGGAGGGCCACGAAGTGGCTT
ATGGGTTTCTCAACCGTGGGAATATTGATGCTTTGAGCTGGATAATTCTTTATTGTGGGGACCGT
GCTTTGAAGGATTTTAAGCATCCCTGGCCACCACCACTAGATGCTGATTAGCACAGGCATGAAAT
ATTCTTCCAACCTTGCCGAAAGTCGCTGGGAGATCATTCCTTCCGATGAGAATACTGGGCCGGGG
CCTGACAGCCTTGGGTCCCAAAGGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAG
AACCCTTCAGCAGCCGGGGC
```

Or, choose a file to upload

Browse...

Set subsequence: (optional)

From: To:

Database:
genome (reference only) 741 sequences

Program:
cross-species megaBLAST: Compare nucleotide sequences for other species to this genome

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options: -G 5 -E 2

Begin Search Clear Input

Get the URL with preset values ? Get URL

BLAST *Basic Local Alignment Search Tool*

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/BLAST/Format Request

Query lcl|31017 (2750 letters)

Database ref_contig

Job title lcl|31017 (2750 letters)

Request ID C3JDWRZW01R View report Show results in a new window

Format

Show Alignment as HTML Advanced View Reset form to defaults

Alignment View Hit Table

Display Graphical Overview Linkout Sequence Retrieval NCBI-gi CDS feature

Masking Character: Lower Case Masking Color: Grey

Limit results Descriptions: 100 Graphical overview: 100 Alignments: 100

Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.

Enter organism name or id--completions will be sugge

Entrez query:

Expect Min: Expect Max:

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ Formatting Results - C3JDWRZW01R [Reformat these Results] [Sign in above to save your search strategy]

Job Title: lcl|31017 (2750 letters)

```
# BLASTN 2.2.17 (Jun-24-2007)
# Query:
# Database: gpipe/10116/ref_contig
# Fields: query id, subject ids, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 5 hits found
31017   gi|62656472|ref|NW_047334.2|Rn10_WGA1854_4      79.33   329     51      2       857     1184    42941092        42941404        1e-54   221
31017   gi|62656472|ref|NW_047334.2|Rn10_WGA1854_4      91.14   158     13      1       1529    1685    42941731        42941888        3e-52   214
31017   gi|62656472|ref|NW_047334.2|Rn10_WGA1854_4      83.61   183     29      1       1947    2129    42942340        42942521        6e-39   169
31017   gi|62656472|ref|NW_047334.2|Rn10_WGA1854_4      79.73   74      15      0       2285    2358    42942641        42942714        8e-05   56.4
31017   gi|34880115|ref|NW_048040.1|RnX_WGA2565_4       86.27   51      7       0       1541    1591    4717108 4717158 2e-05   58.4
```

GreenGene Gene Prediction Analysis Tool

Width of Display 800 | Clear All Exons | Add Exon | Delete Exon

DNA
Genscan
GeneMark
RefSeq Proteins
RepeatMasker
Human ESTs
ProScan
Mouse Genomic
Rat Genomic

Sequence of Selected Feature: GGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC

Sequence of cDNA Assembly

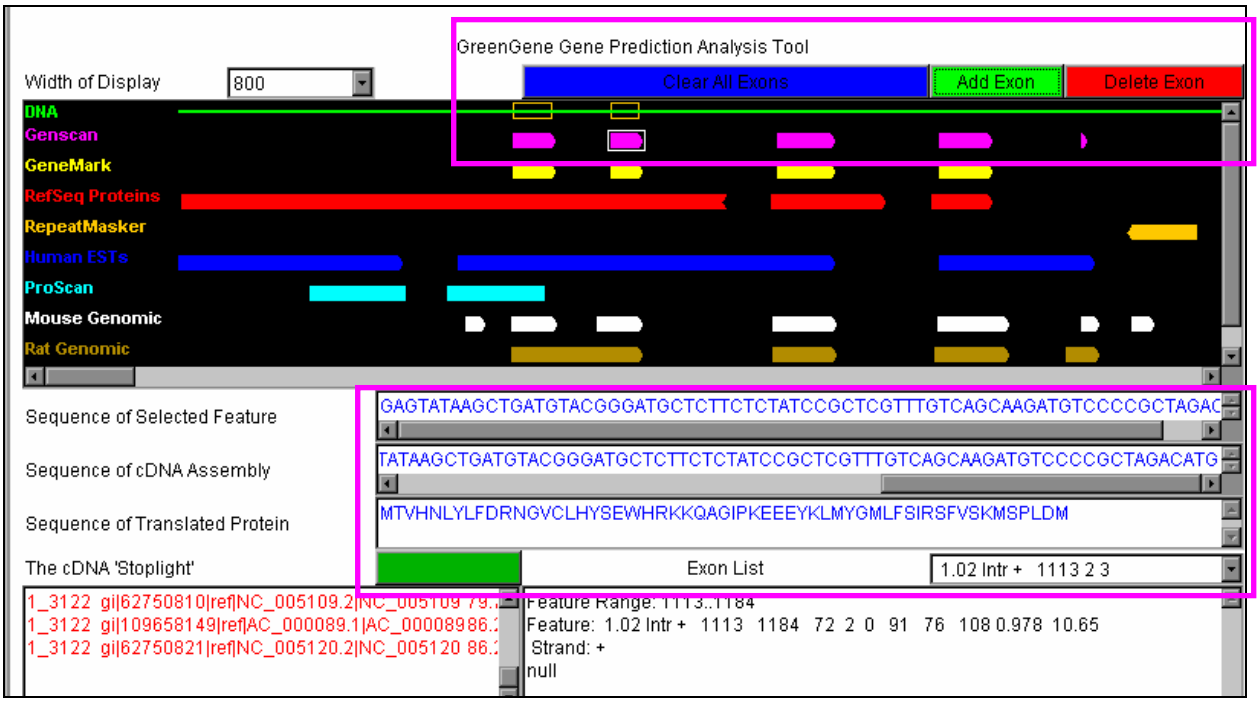
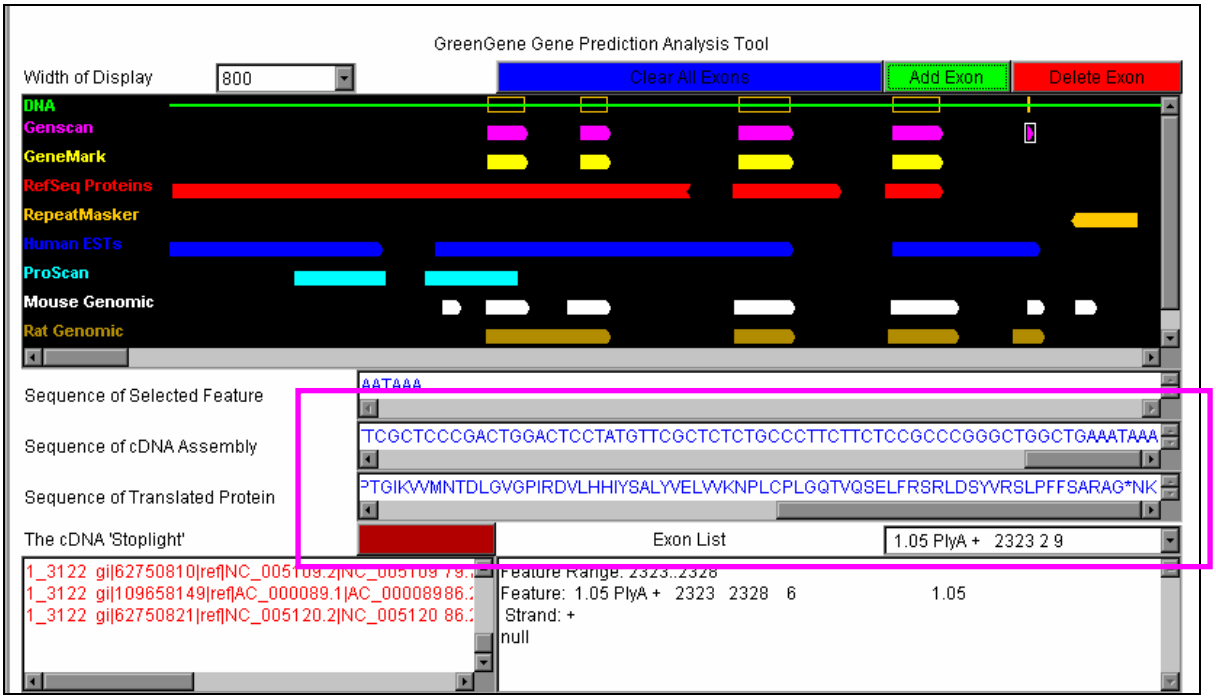
Sequence of Translated Protein

The cDNA 'Stoplight' | Exon List

```
1_3122  gi|62750810|ref|NC_005109.2|NC_005109 79.
1_3122  gi|109658149|ref|AC_000089.1|AC_00008986.
1_3122  gi|62750821|ref|NC_005120.2|NC_005120 86.
```

Feature Range: 1..2750
Feature: null Strand: null
null

Input Area | Tools | Message Console

Run Tool | Clear Input | Rat Genomic (X-Species Megablast) | Import Results

Create a custom coding sequence and the corresponding protein sequence: