
Web-based tools for Bioinformatics; A (free) introduction to (freely available) NCBI, MUSC and World-wide.

When and Where---Wednesdays 1-2pm Room 438 Library Admin Building Beginning September 10, 2003.

Overview Human Genome Resources October 22, 2003

The human genome is largely complete with gaps remaining near chromosome tips and near centromeres. The NCBI process for assembling the contigs which comprise this information is described [here](#). There are a variety of [genetic and physical maps](#) of the human genome which all represent the same thing but because they were measured differently, the alignment of the various maps must be interpreted with a bit of caution. You can access this information from various entry points. The [Genome View](#) offers a scaled view of the 23 chromosomes and the mitochondrial genome. You can [BLAST the human genome](#) or a particular chromosome. The [NCBI Site Map](#) is a useful portal as well. My personal favorite is the [LocusLink](#) portal but the [ENTREZ](#) portal works as well. Another nice tool is the [Human-Mouse Homology Map](#). We will look in later weeks specifically at the ENSEMBL, UCSC and UC Berkeley portals but today we will focus on the NCBI resources.

Introduction/Scope

A worked Example

Continuing with the caspase example let's connect to [LocusLink](#) and search for human "caspase 7".

This brings up the MapViewer window more or less centered on the location of the caspase 7 gene. Note on the left the zoom control and the chromosome ideogram with the position of the screen view on the ideogram. Also note the numerous links to the right of the caspase 7 entry. Note the orientation arrow which represents the 5' to 3' orientation of the caspase 7 gene.

Address [http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?ORG=hum&query=CASP7%5B5YM%5D&CHR=10&MAPS=cntg\[108377536](http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?ORG=hum&query=CASP7%5B5YM%5D&CHR=10&MAPS=cntg[108377536)

MapViewer Home
 Map Viewer Help
 Human Maps Help
 FTP

Data As Table View
Maps & Options
 Compress Map

Region Shown:

out
 zoom
 in

10p15
 10p14
 10p13
 10p12
 10p11+2
 10p11+1
 10p11+2
 10q21
 10q22
 10q23
 10q24
 10q25
 10q26

default
 master

114.8M
 114.9M
 115M
 115.1M
 115.2M
 115.3M
 115.4M
 115.5M

HABP2 ↓ [OMIM](#) [sv](#) [pr](#) [dl](#) [ev](#) [mm](#) [hr](#)

NRAP ↑ [OMIM](#) [sv](#) [pr](#) [dl](#) [ev](#) [mm](#) [hr](#)

CASP7 ↓ [OMIM](#) [sv](#) [pr](#) [dl](#) [ev](#) [mm](#) [hr](#)

FLJ23537 ↓ [sv](#) [pr](#) [dl](#) [ev](#) [mm](#)

DCLRE1A ↑ [sv](#) [pr](#) [dl](#) [ev](#) [mm](#) [hr](#)

LOC374354 ↓ [sv](#) [pr](#) [dl](#) [ev](#) [mm](#)

FLJ20147 ↓ [sv](#) [pr](#) [dl](#) [ev](#) [mm](#)

ADRB1 ↓ [OMIM](#) [sv](#) [pr](#) [dl](#) [ev](#) [mm](#) [hr](#)

FLJ10188 ↑ [sv](#) [pr](#) [dl](#) [ev](#) [mm](#) [hr](#)

[http://www.ncbi.nlm.nih.gov/HomoloGene/homolquery.cgi?TEXT=4892\[loc\]&TAXID=9606](http://www.ncbi.nlm.nih.gov/HomoloGene/homolquery.cgi?TEXT=4892[loc]&TAXID=9606)

At the top of the MapViewer you can see the build number, chromosome number being viewed, the current maps being viewed and the "Maps & Options" button. This button launches another window where various additional maps can be added or removed from the current view.

Address <http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?CHR=10&QUERY=CASP7%5BSYM%5D&MAPS=cntg%5B115000000.000000>

Homo sapiens Map View [BLAST The Human Genome](#)

Build 34 Version 1

Chromosome: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) **[10]** [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#)
[21](#) [22](#) [X](#) [Y](#)

Query: CASP7[SYM] [\[clear\]](#)

Master Map: Genes On Sequence **Maps & Options**

Total Genes On Chromosome: 932 [13 not localized]

Region Displayed: 115,100K-115,160K bp

[Download/View Sequence/Evidence](#)

Genes Labeled: 1 Total Genes in Region: 1

Contig	Clone	HsUniG	Gen...	Symbol	LinkOut	E
115100K						
115101K						
115102K						
115103K						
115104K						
115105K						
115106K						
115107K						
115108K						
115109K						
115110K						
115111K						
115112K						
115113K						
115114K						
115115K						
115116K						
115117K						
115118K						
115119K						
115120K						
115121K						
115122K						
115123K						
115124K						
115125K						
115126K						
115127K						
115128K						
115129K						
115130K						
115131K						
115132K						
115133K						
115134K						

115130K CASP7 + [OMIM](#) [sv](#) [pr](#) [dl](#) [ev](#) [mm](#) [hm](#) C

Here is the Maps & Options window showing the currently displayed maps and the "add/ remove" buttons.

Query: CASP7[SYM] [\[clear\]](#)

Master Map: Genes On Sequence **Maps & Options**

Total Genes On Chromosome: 932 [[13 not localized](#)]

Region Displayed: 115,100K-115,160K bp

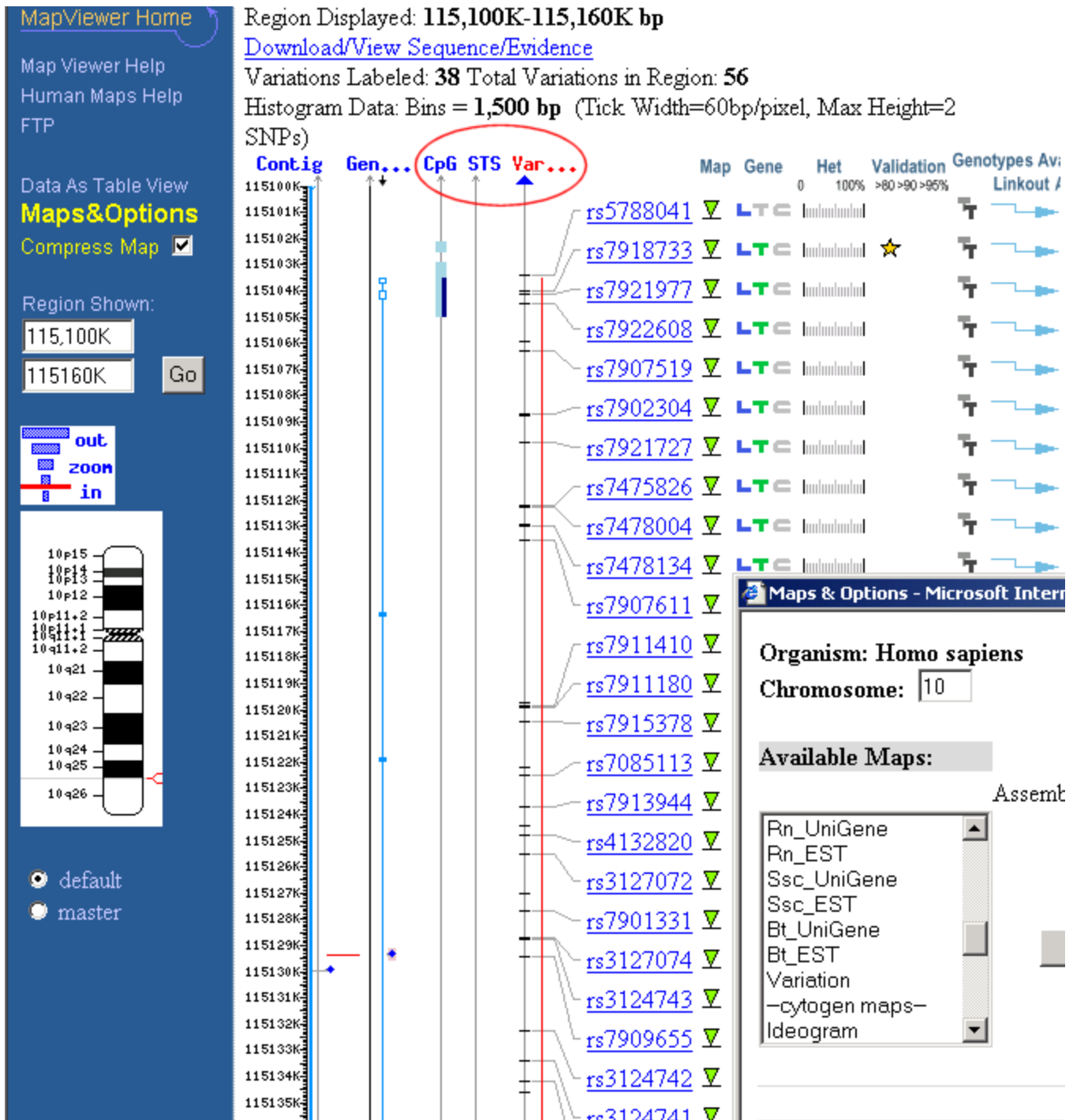
[Download/View Sequence/Evidence](#)

Genes Labeled: 1 Total Genes in Region: 1

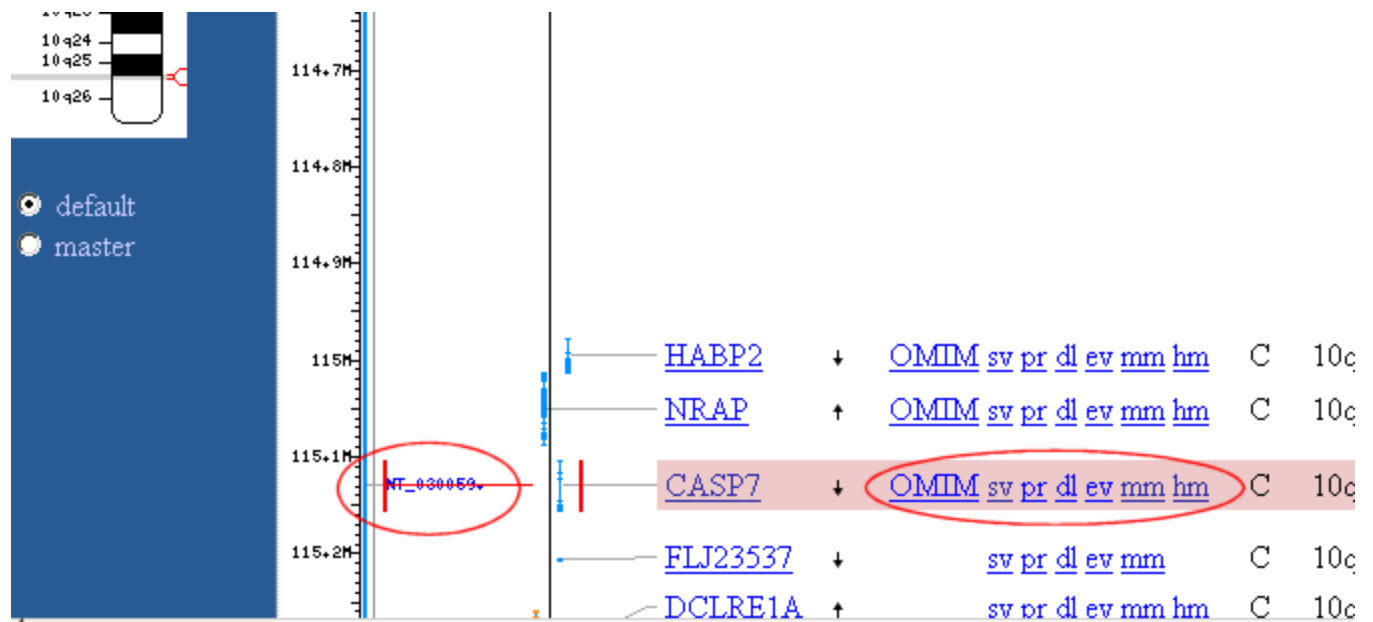
Contig **Clone** **HsUniG** **Gen...** Symb

The image displays a genomic map of the CASP7 gene region on chromosome 10. The map includes a vertical scale from 115,100K to 115,130K bp. Several tracks are visible: 'out' (blue), 'zoon' (red), and 'in' (black). A 'Maps & Options' dialog box is open on the right, showing 'Organism: Homo sapiens', 'Chromosome: 10', and 'Region: 115,100K-115,160K bp'. The dialog lists 'Available Maps' such as 'Ab initio', 'Assembly', 'BES_Clone', 'Clone', 'Contig', 'Component', 'CpG Island', and 'dbSNP haplotype'. The 'More Options' section includes 'Show Connections' (unchecked), 'Verbose M' (checked), and 'Page Length: 40'. The 'Thumbnail View' is set to 'default (ideogram)'. At the bottom of the map, there are links for 'CASP7', 'OMIM', 'sv', 'pr', 'dl', 'ev', 'mm', and 'hm'.

Here is one more image of a more complex map display. This shows some recorded variation for this CASPASE gene. We will return to this topic of variation in a later session of this series.

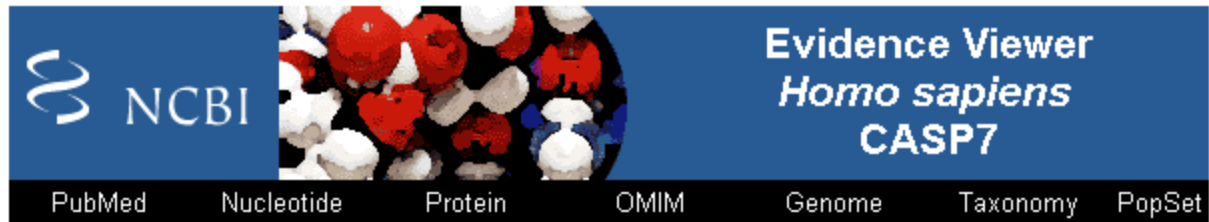


In this next image we can see that for CASP7 there are links to REFSEQ, OMIM, sequence viewer(sv), protein (pr), download (dl), evidence viewer (ev), model maker (mm), and homology (hm).



sequence viewer, evidence viewer and model maker are complex tools, while the others are more straightforward. Here is a screen shot of the sequence viewer.

Below is a screen shot of the Evidence viewer:



NCBI Evidence Viewer
Homo sapiens
CASP7

PubMed Nucleotide Protein OMIM Genome Taxonomy PopSet

Key for display of mRNAs aligning in this region:

[MapView](#)
[Evidence Viewer Help](#)

- Genomic sequence (C)
 - model exons, single (M) ■ mRNA exons, single (G, R)
 - model exons, [overlapping](#) (M) ■ mRNA exons, [overlapping](#) (G, R)
- C = contig, M = model mRNA; R = RefSeq mRNA; G = GenBank mRNA
R = new since last genome build; **R** = updated since last genome build

EST density key (E):

- 1 EST ■ 2-5 ESTs ■ 6-20 ESTs
- 21-99 ESTs ■ >100 ESTs

10 exons and 1 gene found in this genomic region spanning 52500 bp.

[View graphic only](#)



Mouse over mismatches, indels and unaligned regions to see their exon number.

Jump to differences in: [Exon 1](#) [Exon 2](#) [Exon 3](#) [Exon 4](#) [Exon 5](#) [Exon 7](#) [Exon 9](#) [Exon 10](#)

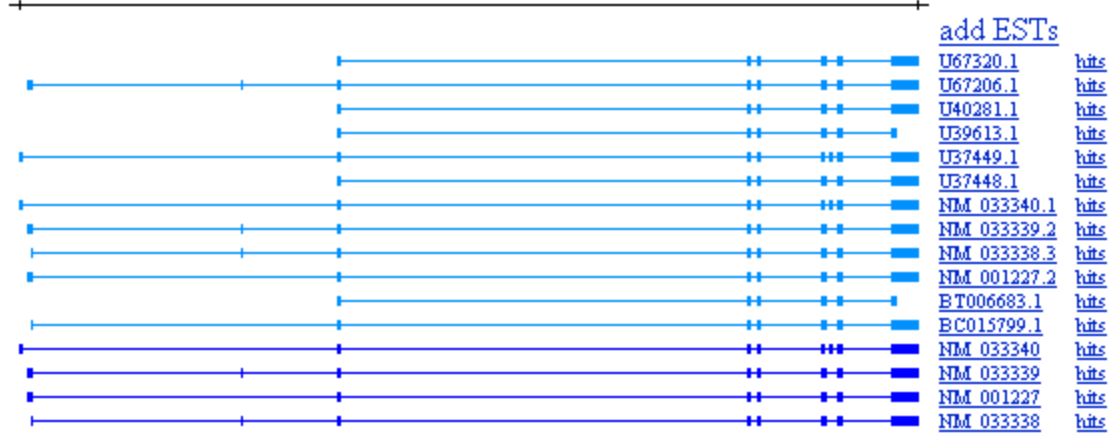
The model maker is the most complex of these viewer in that it allows interactive manipulation of the available data. Here's a basic screen shot:

Model Maker *(Make Your Own Model by selecting an evidence exon "set" and/or add/remove individual putative exons for inclusion in your model)*

[help](#) [legend](#)

Evidence:

34187489<<< [NT_030059.11](#) [mv sv](#) >>>34239188 [change strand](#)
[ev seq](#)



Putative exons (graphic view):



Your model:

[clear](#)

[ORF Finder](#)
[Save](#)

Frame1, ORF= Frame2, ORF= Frame3, ORF=

Putative exons (table view):

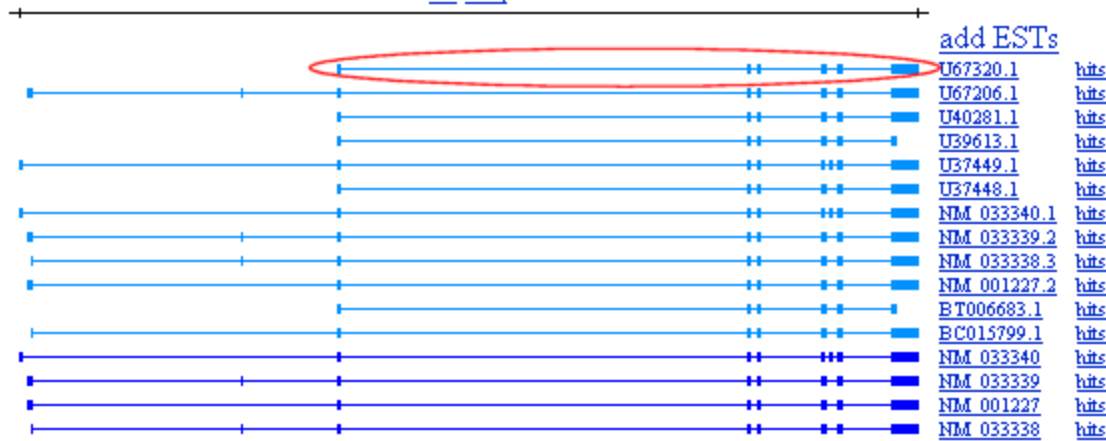
[custom exons](#) [intron bases](#):

<input type="checkbox"/>	1	CA GCA	34187489-34187634	GGG GT =>	5
<input type="checkbox"/>	2	34187954 CT CCC...GA GAC	34188225-34188263	ACG GT =>	3 or 4 or 5
<input type="checkbox"/>	3	2 <= AG ATG	34200249-34200355	CAG GT =>	5
<input type="checkbox"/>	4	2 <= AG GCA	34200282-34200355	CAG GT =>	5
<input type="checkbox"/>	5	1 or 2 or 3 or 4 <= AG ATG	34205779-34205888	CAG GT =>	6
<input type="checkbox"/>	6	5 <= AG TAA	34229317-34229453	CAG GT =>	7
<input type="checkbox"/>	7	6 <= AG GTA	34229936-34230064	AAG GT =>	8 or 9
<input type="checkbox"/>	8	7 <= AG CTT	34233647-34233714	AAT GT =>	10
<input type="checkbox"/>	9	7 <= AG CTT	34233647-34233822	CAG GT =>	11
<input type="checkbox"/>	10	8 <= AG ATG	34234102-34234175	CAG GT =>	11

In the image below I have selected a set of exons and the model maker has collected the appropriate sequence, examined it with ORF finder and displayed the translations of all three frames.

Evidence:

34187489<<< [NT 030059.11](#) [mv sv](#) >>>34239188 [change strand](#)
[ev seq](#)



Putative exons (graphic view):



Your model:

U67320.1, U40281.1, U39613.1, U37448.1, BT006683.1 [clear](#)

```

ATGGCAGATGATCAGGGCTGTATTGAAAGAGCAGGGGGTTGAGGATTCAGCAAATGAAAGAT
TCAGTGGATGCTAAGCCAGACCGGTCCTCGTTTGTACCGTCCCTCTTCAGTAAGAAGAAG
AAAAATGTCACCATGCGATCCATCAAGACCACCCGGGACCGAGTGCCTACATATCAGTAC
AATCATGAATTTTAAAAAGCTGGGCAAATGCATCATAATAAAACAACAAGAACTTTGATAAA

```

[ORF Finder](#)
[Save](#)

Frame1, ORF= Frame2, ORF= Frame3, ORF=

<input type="text" value="MADDQGCIEEQGVEDSANED"/>	<input type="text" value="wqmiravlkserglriqcmki"/>	<input type="text" value="gr*sgly*ragg*gfsk*rf"/>
<input type="text" value="SVDAKPDRSSFVPSLFSSKKK"/>	<input type="text" value="qwmlsqtgprlyrpsvrrr"/>	<input type="text" value="sgc*arplvctvplq*eee"/>
<input type="text" value="KNVTMRSIKTTRDRVPTYQY"/>	<input type="text" value="kmspcdpsrppgteclhist"/>	<input type="text" value="kchhahqdhpgpsayisvq"/>
<input type="text" value="NNFEEKLGKCI I INNKNEFDK"/>	<input type="text" value="t*ilkswanas**ttrtlik"/>	<input type="text" value="hef*kagcmhnhkqqel*s"/>

Putative exons (table view):

[custom exons](#) [intron bases:](#)

<input type="checkbox"/>	1		CA GCA	34187489-34187634	GGG GT=>	5
<input type="checkbox"/>	2	34187954	CT CCC...GA GAC	34188225-34188263	ACG GT=>	3 or 4 or 5
<input type="checkbox"/>	3		2 <= AG ATG	34200249-34200355	CAG GT=>	5
<input type="checkbox"/>	4		2 <= AG GCA	34200282-34200355	CAG GT=>	5
<input checked="" type="checkbox"/>	5	1 or 2 or 3 or 4	<= AG ATG	34205779-34205888	CAG GT=>	6
<input checked="" type="checkbox"/>	6		5 <= AG TAA	34229317-34229453	CAG GT=>	7
<input checked="" type="checkbox"/>	7		6 <= AG GTA	34229936-34230064	AAG GT=>	8 or 9
<input type="checkbox"/>	8		7 <= AG CTT	34233647-34233714	AAT GT=>	10
<input checked="" type="checkbox"/>	9		7 <= AG CTT	34233647-34233822	CAG GT=>	11
<input type="checkbox"/>	10		8 <= AG ATG	34234102-34234175	CAG GT=>	11
<input checked="" type="checkbox"/>	11	9 or 10	<= AG GCT	34234590-34234719	CAG GT=>	12
<input checked="" type="checkbox"/>	12		11 <= AG GCT	34237596-34239181	CTT AATTCCT AA	34239188

Additional Human Genome Resources.

UniGene

UniGene is a database of clusters(clusters created via BLAST algorithm means)of transcription product sequences. Sequences and sequence fragments that are determined via BLAST to correspond to a given gene make up the cluster. ESTs for instance are recorded in UniGene database. To make use of UniGene, for example, you could search for

"caspase 7 AND human [orgn]".

From the UniGene page you can download the sequences or use the [MapViewer to visualize them](#).

Expressed Sequence Tags and **dbEST**

High-Throughput Genomic Sequences **HTGS**

HTGs (High-Throughput Genomic Sequences) are "unfinished" DNA sequences generated by the high-throughput sequencing centers. The idea is to make this sequence publicly available as quickly as possible.

- HTG records consist of first read sequence data generated from a single cosmid, BAC, YAC, or P1 clone.
- HTG sequence is greater than 2kb in size.
- HTG sequence contains gaps
- HTG sequence is assigned a single accession number
- HTG records include a clear indication of unfinished status

HTG entries are characterized by a phase number 0-3.

- phase 0 single pass clone no contigs
- phase 1 contigs which may be unfinished,unordered,unoriented and contain gaps.
- phase 2 contigs that are unfinished but ordered, oriented with or without gaps.
- phase 3 conigs that are finished, no gaps and may or may not have annotations.

Sequence Tagged Sites **dbSTS**

Sequenced Tagged Sites (STSs) are short (about 200-500 bp) sequences that are unique in a genome. The exact location and order of the bases of the sequence must be known. The majority of STSs consist of a pair of primers that will amplify a specific, unique piece of DNA from the chromosome or genome. ESTS can also be used as STSs.

- STSs can be specifically detected by PCR in the presence of all other genomic sequences
- STSs define a specific position on the physical map of the genome
- STSs are used as mapping reagents

UniSTS and **e-PCR**

UniSTS is a unified, non-redundant view of sequence tagged sites (STSs) which integrates marker and mapping data from a variety of public resources, including, but not limited to, dbSTS. Other sources include the GDB (Genome Database), the Genethon genetic map, the Marshfield genetic map, and the Whitehead RH map.

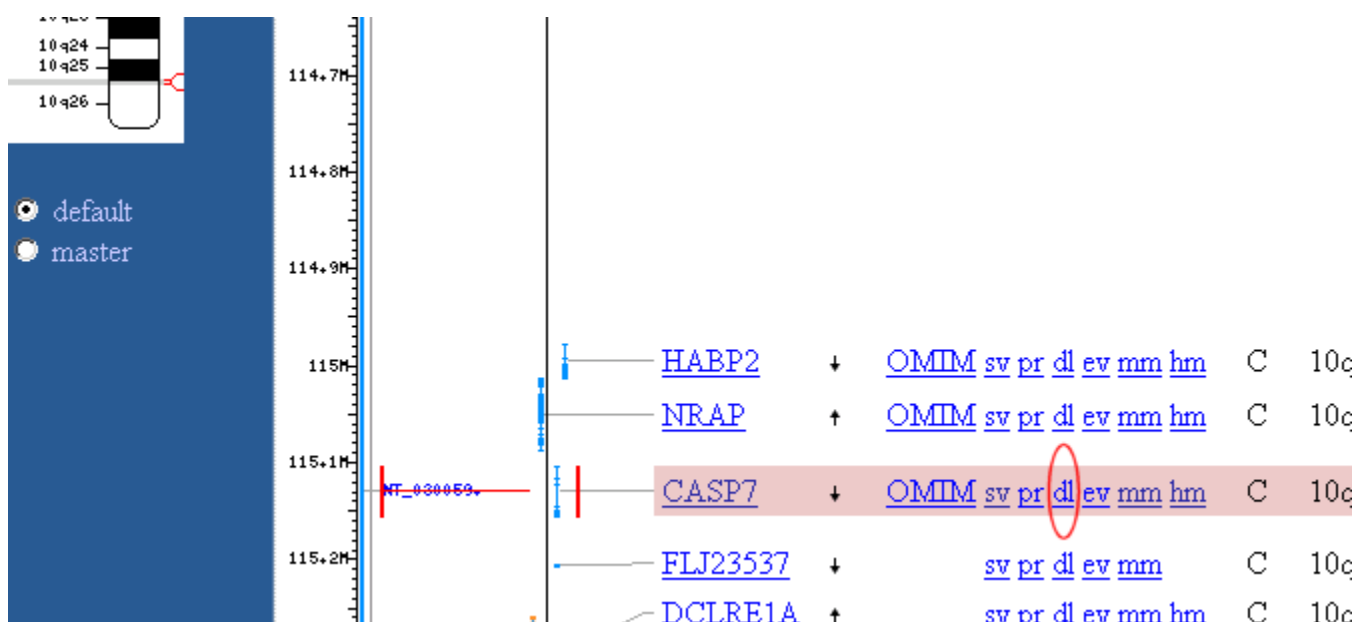
For each marker, the [UniSTS record displays \(in this case the query was "caspase 7"](#)

- the primer sequences
- product size
- mapping information
- cross references to LocusLink, dbSNP, RHdb, GDB, MGD, and the Entrez Map Viewer
- a list of GenBank and RefSeq records that contain the primer sequences as determined by Electronic PCR (e-PCR)


e-PCR is a tool that searches the data in NCBI's UniSTS for mapped (known) STSs in a query sequence in order to determine a possible map location. E-PCR finds STSs in DNA sequences by identifying subsequences that match the marker's PCR primer pairs. The subsequences must have the correct order, orientation, and spacing to amplify a PCR product of the correct molecular weight.

Sample Questions/Data

How can I find and download the 1000bp which flank the caspase 7 gene? Go to LocusLink, search for Caspase 7. On the caspase 7 LocusLink page click on the MAP icon or further down in the entry click on "mv" to see something like the following image. Then click on the download (dl) link.



On the DownLoad page simply adjust the boundaries by the desired amount and save the result to your PC.

Address  http://www.ncbi.nlm.nih.gov/mapview/seq_reg.cgi?chr=10&from=115103550&to=115155249

Homo sapiens Genome
Region to retrieve (in chromosome coordinates):
Chromosome: Strand:
from: adjust by:
to: adjust by:

Sequence Format:

This chromosome region corresponds to the contig region(s):

Contig	start	stop	strand	
NT_030059.11	34187489	34239188	+	Display Save to Disk View Evidence ModelMaker



Created by ESH 8-18-2003; updated 10-21-2003 11:30

[email to Starr about this page](#)