

Sequence Databases

MASCOT

 **MATRIX
SCIENCE**

The collage features the following elements:

- NCBI Data & Software**: A screenshot of the NCBI website's 'Data & Software' section, showing links to databases, downloads, and tools.
- UniProt**: The UniProt logo, a central hub for protein sequence data.
- WormBase**: The WormBase logo, a database for the genome and gene expression of the nematode *Caenorhabditis elegans*.
- gpm**: The gpm logo, likely representing the Genomic Protein Map project.
- TrEMBL**: The TrEMBL logo, a database of protein sequences that have not yet been manually reviewed.
- Proteomes**: A logo with icons representing different proteomes.
- Zebrafish**: The Zebrafish logo, representing the Zebrafish Genome Project.
- swissprot**: The swissprot logo, a database of protein sequences that have been manually reviewed.
- nextprot**: The nextprot logo, a database of protein sequences that have been manually reviewed.
- NCBI**: The NCBI logo, the National Center for Biotechnology Information.
- EMBL-EBI**: The EMBL-EBI logo, the European Molecular Biology Laboratory-European Bioinformatics Institute.
- MASCOT**: The MASCOT logo, a search engine for protein sequences.
- Matrix Science**: The Matrix Science logo, a company that provides protein search engines.
- EMBL-EBI Services**: A screenshot of the EMBL-EBI 'Services' page, showing various tools and data resources.

When you install Mascot, it includes a copy of the Swiss-Prot protein database. However, it is almost certain that you and your colleagues will want to search other databases as well. There are very many to choose from, and Mascot allows you to have as many databases on-line for searching as you wish.

Matrix Science doesn't supply sequence databases. Most databases are public domain, and there are a few sites that provide comprehensive database repositories. Two of the best known are NCBI and EBI. Here, you can download nr, GenBank, Swiss-Prot, EMBL, TrEMBL, etc.

For specialised databases, such as individual genomes, you may have to track down the FTP site of the group that is doing the sequencing.

Sequence Databases

Swiss-Prot (~564,000 entries)

- High quality, non-redundant; ideal for PMF & some MS/MS

UniProt proteome database (size varies by species)

- >300K proteomes of which 18K are reference proteomes
- Quality varies depending on popularity of species

NCBIprot, UniRef100 (NCBIprot ~340,000,000 entries)

- Comprehensive, non-identical

EST databases (>400,000,000 entries in translation)

- Very large and very redundant
- Not suitable for PMF

Sequences from a single genome

- Not suitable for PMF

There are a huge number of database, and often it is not clear which is the appropriate one to choose for a search.

SwissProt is acknowledged to be the best annotated database, and is non-redundant, making it an ideal choice for PMF searches, where the loss of one or two peptides is not a concern. SwissProt is also a good choice for MS/MS of a well characterised organism, such as human or mouse or yeast.

UniProt proteome database for the species of interest are an excellent database to choose especially if the species is of research importance, Human, Rat, Mouse, E. Coli etc as they will be well annotated and compressive. For less commonly analysed species they can still be a good resource that is a smaller database to search than say all of green plants in NCBIprot. The proteomes are based on the translation of a completely sequenced genome and will normally include sequences that derive from extra-chromosomal elements such as plasmids or organellar genomes in organisms. Some proteomes may also include protein sequences based on high quality cDNAs. The raw sequence data comes from translations of genome sequence submissions to the International Nucleotide Sequence Database Consortium ([INSDC](#)). Proteomes with a Benchmarking Universal Single-Copy Orthologs (BUSCO) complete score above 95% considered good.

The comprehensive, non-identical databases are a good choice for MS/MS searching if you don't want to miss any matches. After NCBI changed the accession number formatting in 2017 the nr database definition is now called NCBIprot on Mascot Server.

NCBIprot and UniRef100 both aim to include explicit representations of all known protein sequences. However, they are huge, over 300 million entries so take a long time to search. Plus, only the best quality data will obtain matches when searching the whole database. There are some non-redundant versions of UniProt100, such as UniRef90 and UniRef50, if you search these databases you may miss some matches.

If the genome of your organism of interest has not been sequenced, it won't be represented in the protein databases, but there may be lots of Expressed Sequence Tags (ESTs). Not advisable for PMF, because many sequences correspond to protein fragments.

Single genome databases can sometimes be useful for MS/MS searches. You will want to include a contaminants database in the search, to ensure spectra from contaminants don't get mis-assigned to the target organism

(Entry counts from mid 2022)

NA Translation

K P I R L T A D L L A E T L Q A R R E W G P I F N I
 S P S D # Q Q I S W Q K L Y K P E E S G G Q Y S T H
 Q A H Q T N Q S R S L G R N S T S Q K R V G A N I Q H
 AAGGCCCTCAGACTAAACAGCAGATCTCTTGGCAGAACTCTACAGCCGGAAGAGAGTGGGGGCCAATATTCACATT
 [299200] [299210] [299220] [299230] [299240] [299250] [299260] [299270]
 TTCGGGTAGTCTGATTGCTCTAGAGAACCGTCTTTGAGATGTTCCGGTCTCTCTCACCCCGGTTATAAGTTGTAA
 A W * V L L L D R P L F E V L W F L T P A L I * C E
 L G D S + C C I E Q C F S + L G S S L P P W Y E V N
 L G M L S V A S R K A S V R C A L L S H P G I N L M

```
Residue: FFLSSSSSY*CC*WLLLLPPPHHQRRRIIIMTTTTNNKKSSRRVVVAAADDEEGGGG
Start: -----M-----
Base 1: TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCAAAAAAAAAAAAAGGGGGGGGGGGGGGGGGGG
Base 2: TTTTCCCAAAAGGGGTTCCTCCCAAAGGGGTTCCTCCCAAAGGGGTTCCTCCCAAAGGGG
Base 3: TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

* = stop

When we search a nucleic acid databases, Mascot always performs a 6 frame translation on the fly. That is, 3 reading frames from the forward strand and 3 reading frames from the complementary strand.

NA Translation

- Mascot translates on the fly in all 6 reading frames
- Translation starts from the beginning of the sequence, not from a start codon
- When a stop codon is encountered, inserts a gap and re-starts translation
- No attempt to resolve codon ambiguity
- Where taxonomy information is available, translation uses the correct genetic code.

The rules for NA translation in Mascot are

Translate the entire sequence, don't look for a start codon to begin

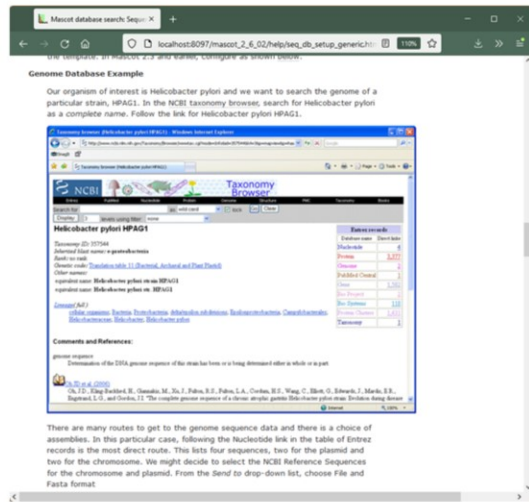
When a stop codon is encountered, leave a gap, and immediately re-start translation

There is no attempt to resolve ambiguous codons. For example, ACX can be translated as Threonine, because the identity of the last base is a don't care. However, this is not done in Mascot.

Finally, all translations use the correct genetic code, as long as the taxonomy is known.

Single Genome Data

Mascot help pages describe how to navigate NCBI web site



MASCOT

: *Sequence Databases*

© 2007-2022 Matrix Science

**MATRIX
SCIENCE**

All the genomes in GenBank are translated into protein sequences in NCBIprot. Usually, this is the simplest option for a Mascot search. But, if you are not confident that the coding sequences and reading frames have been identified correctly, or you are looking for something unusual, you might wish to search the genomic DNA directly. The Mascot help page for a generic database describes how to locate and download different types of sequence data, including genomic DNA -

http://www.matrixscience.com/help/seq_db_setup_generic.html

Single Genome Data

Assembled genomes

- Searching a database of one, (or a few), very long sequences is possible, but:
 - Mascot reports will be unwieldy
 - Memory inefficient
 - Better to split the sequence into segments
 - Small overlaps to ensure no peptide lost
 - Maintain frame numbering
- www.matrixscience.com/downloads/splitter.pl.gz

Assembled genomes are not ideal for a Mascot search, because it would make the reports too unwieldy.

The longest human chromosome is chromosome 1 with 285 million base pairs

We don't know of any tools for reviewing the results which can handle 250 Mbp sequences.

Mascot requires a significant memory overhead to manipulate such long sequences, which means that unless you have a very large amount of RAM, the search is going to be using virtual memory ... i.e. swapping out to disk ... and run relatively slowly.

So, we recommend working with contigs or just chopping the chromosomes into more manageable lengths.

For efficient searching and reporting, the genomic DNA needs to be chopped into shorter sequences, with small overlaps to ensure no peptides are lost because they span a boundary. This is not a completely trivial task if you want to maintain the original forward and reverse frame numbering from chunk to chunk. A simple perl utility to split a long sequence can be downloaded from the Matrix Science web site.

MASCOT Search Results

User: [email]@matrixscience.com
 Search file: PRG2008 Unlabeled Mouse
 MS data file: D:\PRG2008.mgf\merged.mgf
 Database: UniProt, M. musculus (20210502) (34,639 sequences, 38,552,995 residues)
 Time/Lamp: 4 May 2022 at 14:02:09 GMT

Search: ☐ All ☐ Non-significant ☐ Unassigned ☐ Help

Not what you expected? Try [flex search summary](#).

Search parameters

Score distribution

Identification statistics for all protein families

Legend

Protein Family Summary

Significance threshold p: 0.01 Max. number of families: 100
 Target FDR (overlaps eq. threshold): 0.01 FDR type: FDR
 Display non-sig. matches: ☐ Min. number of sig. unique sequences: 1
 Show Fasta scores: ☐ Dendrogram cut at: 0
 Preferred taxonomy: All entries

Identifiability and FDR (reversed protein sequences)

Proteins (148): [Sort By](#) [Unassigned \(2020\)](#) [1 protein](#)

Protein families 1-10 (out of 431)

10 per page: 1 2 3 4 5 6 Next Request all Collapse all Clear

Accession	Protein	Description
1 F20029	1 Q64508	1223 Endoplasmic reticulum chaperone GRP 94Hsc protein (Hs) (Hs)
2 P63017	2 P56655	227 Heat shock cognate 70 kDa protein (Hs) (Hs)
3 P56656	3 P56656	227 Heat shock cognate 70 kDa protein (Hs) (Hs)
4 Q64508	4 Q64508	227 Heat shock cognate 70 kDa protein (Hs) (Hs)
5 P56655	5 P56655	227 Heat shock cognate 70 kDa protein (Hs) (Hs)
6 P56656	6 P56656	227 Heat shock cognate 70 kDa protein (Hs) (Hs)
7 Q64508	7 Q64508	227 Heat shock cognate 70 kDa protein (Hs) (Hs)
8 Q64508	8 Q64508	227 Heat shock cognate 70 kDa protein (Hs) (Hs)
9 P56655	9 P56655	227 Heat shock cognate 70 kDa protein (Hs) (Hs)
10 P56656	10 P56656	227 Heat shock cognate 70 kDa protein (Hs) (Hs)

MASCOT : Sequence Databases © 2007-2022 Matrix Science **MATRIX SCIENCE**

To illustrate the features of the different types of database, we first searched a very small dataset of 33 thousand MS/MS spectra against a protein database, the Uniprot complete mouse proteome. There are significant matches to some 431 Mouse proteins.

Proteins (420)
Report Builder
Unassigned (21841)
[Log out](#)

Protein families 1 - 10 (out of 346)

15 per page
1
2
3
4
5
6
- 35
Next
Expand all
Collapse all
Find
Clear

p1

1 BY012418
2 W91084

700 89103418.1 Mus musculus lung A2B-0356 LLC cDNA, 330bp full-length enriched library ...
659 W91084.1 mg36132-1 Scarae mouse embryo NBE13.5 14.5 Mus musculus cDNA clone ...

p2

AA002359
CX120581

617 AA002289.1 mg43616-1 Scarae mouse embryo NBE13.5 14.5 Mus musculus cDNA cl...
477 CX120581.1 WNA02767 Embryonic day 10 Mouse Pancreas Amplified cDNA library Mus ...

p3

1 B1145268
2 B1221323

430 B1145268.1 B1221323.1 NCL_C04F_109 Mus musculus cDNA clone IMAGE:5051708 F...
147 B1221323.1 B1221323.1 NCL_C04F_109 Mus musculus cDNA clone IMAGE:5051708 F...

p4

1 AW012478
2 AA002379
3 AA002379

383 AW012478.1 u03611.1 Sugeno mouse liver mla Mus musculus cDNA clone IMAGE:25...
64 AA002379.1 m03611.1 Sugeno mouse liver mla Mus musculus cDNA clone IMAGE...
201 AA002379.1 u03611.1 Sugeno mouse liver mla Mus musculus cDNA clone IMAGE:182...
63 AA002379.1 u03611.1 Sugeno mouse liver mla Mus musculus cDNA clone IMAGE:188...
122 AA002379.1 u03611.1 Sugeno mouse liver mla Mus musculus cDNA clone IMAGE:188...
140 AA002379.1 u03611.1 Sugeno mouse liver mla Mus musculus cDNA clone IMAGE:145...
171 AA002379.1 m03611.1 Scarae mouse liver mla Mus musculus cDNA clone IMAGE:189007...

p5

AA000970
B1220869
B1227647
CK624204

352 AA000970.1 mg50101-1 Scarae mouse embryo NBE13.5 14.5 Mus musculus cDNA cl...
347 B1220869.1 B1227647.1 NCL_C04F_109 Mus musculus cDNA clone IMAGE:510281 F...
278 B1227647.1 B1227647.1 NCL_C04F_109 Mus musculus cDNA clone IMAGE:510281 F...
137 CK624204.1 m136054.1 Mouse RPE (choroid, unamplified) m136054 Mus musculus cDNA cl...

p6

1 AW413050
2 AA189729

341 AW413050.1 u031604.1 Sugeno mouse liver mla Mus musculus cDNA clone IMAGE:29...
68 AA189729.1 m136054.1 Scarae mouse lymph node NBE13.5 Mus musculus cDNA clone 1...

p7

1 CF169338
2 A0036073
3 B0064785
4 D0057345

339 CF169338.1 B0064785.1 NIA Mouse Random Kidney cDNA Library (Long 1) Mus muscu...
83 A0036073.1 Mus musculus brain cDNA clone M1C3-7128 F' end...
338 B0064785.1 B0064785.1 NCL_C04F_109 Mus musculus cDNA clone IMAGE:4934718...
297 D0057345.1 H000004_03_001.1 F0 H000004 Mus musculus cDNA clone H000004_03...

p8

1 CK232350
2 BQ091956
3 BE307099

335 CK232350.1 BQ091956.1 Mus musculus hematopoietic B0-HPC8 cDNA library Mus muscu...
170 BQ091956.1 BQ091956.1 NCL_C04F_109 Mus musculus cDNA clone IMAGE:4934718...
106 BE307099.1 B0064785.1 NCL_C04F_109 Mus musculus cDNA clone IMAGE:4934718...

15 per page
1
2
3
4
5
6
- 35
Next
Expand all
Collapse all

MASCOT : Sequence Databases © 2007-2022 Matrix Science

With Mus_EST, we obtained a very similar set of peptide matches. However, look at the hit-list. Unlike the protein database search, it doesn't immediately communicate which proteins have been found. I'll return to this issue later.

	Score	Mass	Matches	Sequences	emPAI	F											
5.1	383	15561	23 (23)	4 (4)	3.58	2	AW012476.1 uc05d11.y1 Sugano mouse liver mla Mus musculus cDNA clone IMAGE:282317 5' siml...										
P53 sameasets of AW012476																	
5.2	201	25379	14 (14)	5 (5)	1.56		A1526761.1 uc42d11.y1 Sugano mouse liver mla Mus musculus cDNA clone IMAGE:192236 5' simlar...										
5.3	171	23932	6 (6)	2 (2)	0.49	1	AA238951.1 uc04d12.y1 Soares mouse lml Mus musculus cDNA clone IMAGE:694057 5' similar to gh...										
P20 sameasets of AA238951																	
5.4	140	33937	17 (17)	5 (5)	1.02	1	A1047293.1 uc04d07.y1 Sugano mouse liver mla Mus musculus cDNA clone IMAGE:1450492 5' siml...										
P2 sameasets of A1047293																	
5.5	122	31131	9 (9)	3 (3)	0.59	6	A1132230.1 uc32d09.x1 Sugano mouse liver mla Mus musculus cDNA clone IMAGE:1482088 5' siml...										
P2 sameasets of A1132230																	
5.6	64	25761	2 (2)	2 (2)	0.45	2	AAB82179.1 uc38d10.y1 Stratagene mouse lung 937302 Mus musculus cDNA clone IMAGE:1277936 5'...										
P2 sameasets of AAB82179																	
5.7	63	20243	6 (6)	2 (2)	0.60	2	A1526994.1 uc01c11.y1 Sugano mouse liver mla Mus musculus cDNA clone IMAGE:1888820 5' simlar...										
P2 sameasets of A1526994																	
▼56 peptide matches (21 non-duplicate, 35 duplicate)																	
Auto-fit to window																	
Query Index	Observed	Mr (exp1)	Mr (calc)	Delta M	Score	Expect	Rank	Q	1	2	3	4	5	6	7	Peptide	
f4447 P3	521.2416	1560.7029	1559.8187	0.8842	0	59	0.00065	P1	0								K.NISQPTWFWSE.A
f4705	525.4566	1573.3479	1572.7654	0.5824	0	71	0.00028	P1	0								K.SALVONDERFAGD.G
f5144 P4	540.3247	1678.6349	1678.5385	0.0964	0	54	0.0091	P1	0								R.SCARGLAR.H
f5615 P2	541.3661	1680.7177	1680.4059	0.3118	0	44	0.007	P1	0								K.DYFPMV.V
f7790	577.9297	1553.8449	1553.4045	0.4404	0	49	0.030	P1	0								R.OFFPMSE.I
f8340	586.6058	1169.9970	1169.5994	0.3976	0	33	0.028	P1	0								R.OFFPMSE.I - Oxidation (M)
f15114 P2	739.5340	1477.0534	1476.8634	0.1900	0	82	7.2e-05	P1	0								K.OTTVITSLSVLR.D
f18138	739.6176	1477.2207	1476.8108	0.4100	0	63	0.0026	P1	0								R.DFIDVYLK.Q
f19443 P2	745.4500	1489.8954	1489.8064	0.0846	0	73	0.00021	P1	0								K.OTTVITSLSVLR.D
f19465 P1	510.6492	1528.9257	1528.7756	0.1500	0	59	0.0019	P1	0								K.SALVONDERFAGD.G
f19473	745.5187	1529.0228	1528.7756	0.2472	0	93	3.3e-06	P1	0								K.SALVONDERFAGD.G
f19851 P2	773.1436	1544.2725	1543.8521	0.4205	0	61	0.00023	P1	0								VGVVVVYLYLR.K
f20247 P7	781.1422	1560.2498	1559.8187	0.4311	0	80	5.3e-06	P1	0								K.NISQPTWFWSE.A
f20594	787.5844	1573.0743	1572.7654	0.3089	0	43	0.040	P1	0								K.SALVONDERFAGD.G
f21390 P6	536.3278	1605.9617	1605.8232	0.1384	1	44	0.014	P1	0								K.SALVONDERFAGD.G
f21414 P1	804.1063	1606.1980	1605.8232	0.3747	1	45	0.014	P1	0								K.SALVONDERFAGD.G
f22532	554.6924	1663.0254	1660.7814	0.2440	0	49	0.014	P1	0								K.SALVONDERFAGD.G
f23907	849.6286	1735.2427	1734.8402	0.4025	0	61	0.0030	P1	0								R.VQERACLVNELA
f25449	614.1602	1639.4589	1638.9402	0.5184	1	42	0.030	P1	0								K.DYFPMVFWSE.H
f26750 P6	645.2435	1932.7686	1932.0772	0.6914	0	50	0.0017	P1	0								K.OTTVITSLSVLR.D
f50244	776.1600	2325.4582	2325.2824	0.1758	0	43	0.023	P1	0								R.FIDLINPLRDEVTSDIK.F
▼36 subsets and intersections (215 subset proteins in total)																	
P8247578																	
Score	370	36187	5.1					B1247578.1 60296039F1 NCL_CGAP_L9 Mus musculus cDNA clone IMAGE:5125617 5' mRNA sequence.									
AA575398	370	16504	5.1					AA575398.1 uc7909.y1 Stratagene mouse diaphragm (493703) Mus musculus cDNA clone IMAGE:987641 5' siml...									
P22 sameasets of AA575398																	
A1097827																	
Score	365	19139	5.1					A1097827.1 uc36d05.y1 Sugano mouse liver mla Mus musculus cDNA clone IMAGE:1482489 5' similar to gh:01767...									
P22 sameasets of A1097827																	
BF234254																	
Score	348	38548	5.1					BF234254.1 602208142F1 NCL_CGAP_L9 Mus musculus cDNA clone IMAGE:4161347 5' mRNA sequence.									
P7 sameasets of BF234254																	
A1529126																	
Score	344	21653	5.1					A1529126.1 uc09f1.y1 Sugano mouse liver mla Mus musculus cDNA clone IMAGE:1888809 5' similar to gh:017674...									
A1785827																	
Score	335	13453	5.1					A1785827.1 uc79d3.y1 Sugano mouse liver mla Mus musculus cDNA clone IMAGE:1888660 5' similar to gh:M6185...									

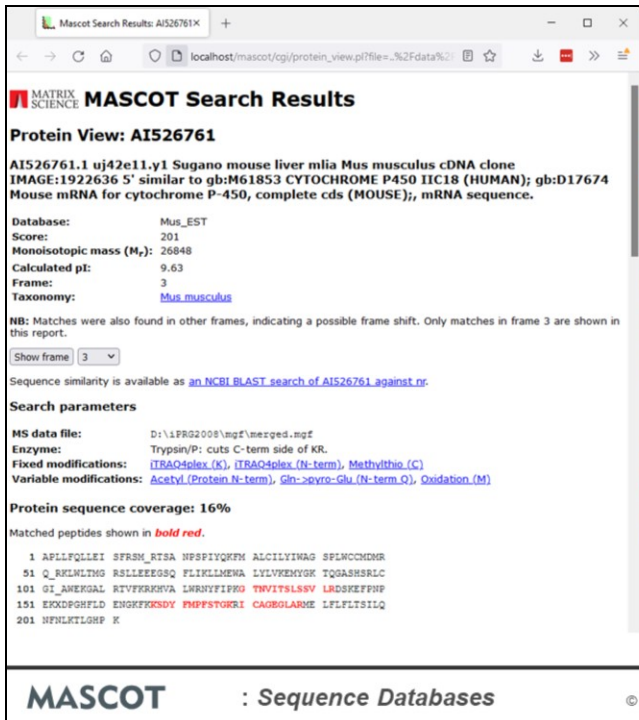
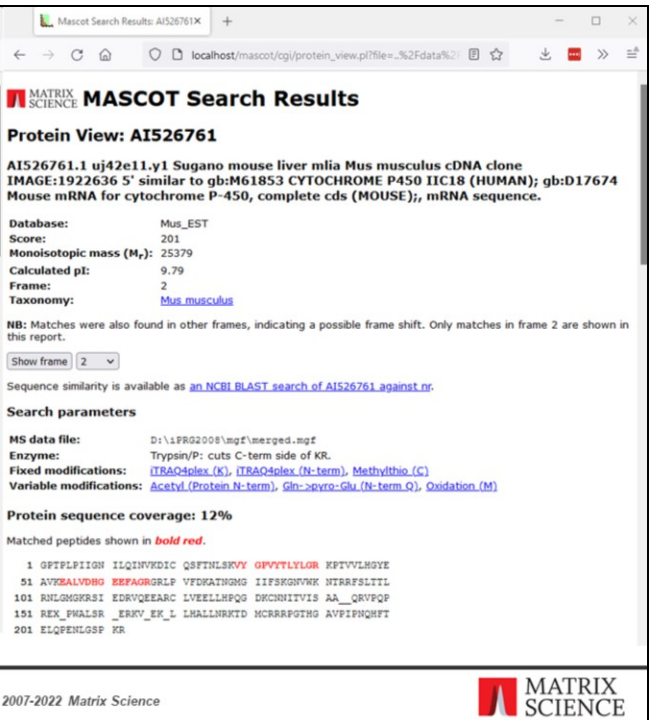
MASCOT

: Sequence Databases

© 2007-2022 Matrix Science



The Protein Family results report from the EST search looks pretty similar to the UniProt search, except that the EST sequences are mostly shorter than full length proteins, so the peptide matches are more scattered. If we click on a protein accession number link

MASCOT Search Results

Protein View: AI526761

AI526761.1 uj42e11.y1 Sugano mouse liver mlaia Mus musculus cDNA clone
IMAGE:1922636 5' similar to gb:M61853 CYTOCHROME P450 IIC18 (HUMAN); gb:D17674
Mouse mRNA for cytochrome P-450, complete cds (MOUSE);, mRNA sequence.

Database: Mus_EST
Score: 201
Monoisotopic mass (M_r): 26848
Calculated pI: 9.63
Frame: 3
Taxonomy: [Mus musculus](#)

NB: Matches were also found in other frames, indicating a possible frame shift. Only matches in frame 3 are shown in this report.

Show frame | 3 |

Sequence similarity is available as [an NCBI BLAST search of AI526761 against nr](#).

Search parameters

MS data file: D:\IPRG2008\mgf\merged.mgf
Enzyme: Trypsin/P: cuts C-term side of KR.
Fixed modifications: [ITRAQ4plex \(K\)](#), [ITRAQ4plex \(N-term\)](#), [Methylation \(C\)](#)
Variable modifications: [Acetyl \(Protein N-term\)](#), [Gln->pyro-Glu \(N-term Q\)](#), [Oxidation \(M\)](#)

Protein sequence coverage: 16%

Matched peptides shown in **bold red**.

```

1  APLFLQLLEI SFRSM_RISA NPSPIYQKFM ALCLLYINAG SFLWCHDNR
51  Q_RKLMLTMS RSLLEERGSQ FLIKILMENA LVLVKNYQK TQGSASRLC
101  Q1_ANEKGAH RTVFERKHVA LNRNYFIFGQ THVITSLSSV LRDSKEFFHP
151  EKXDPQHFLD ENKFKQSDY PMPFSTQKRI CAGGLARKE LFLFLTSILQ
201  NFNKLTLQHP K

```

MASCOT Search Results

Protein View: AI526761

AI526761.1 uj42e11.y1 Sugano mouse liver mlaia Mus musculus cDNA clone
IMAGE:1922636 5' similar to gb:M61853 CYTOCHROME P450 IIC18 (HUMAN); gb:D17674
Mouse mRNA for cytochrome P-450, complete cds (MOUSE);, mRNA sequence.

Database: Mus_EST
Score: 201
Monoisotopic mass (M_r): 25379
Calculated pI: 9.79
Frame: 2
Taxonomy: [Mus musculus](#)

NB: Matches were also found in other frames, indicating a possible frame shift. Only matches in frame 2 are shown in this report.

Show frame | 2 |

Sequence similarity is available as [an NCBI BLAST search of AI526761 against nr](#).

Search parameters

MS data file: D:\IPRG2008\mgf\merged.mgf
Enzyme: Trypsin/P: cuts C-term side of KR.
Fixed modifications: [ITRAQ4plex \(K\)](#), [ITRAQ4plex \(N-term\)](#), [Methylation \(C\)](#)
Variable modifications: [Acetyl \(Protein N-term\)](#), [Gln->pyro-Glu \(N-term Q\)](#), [Oxidation \(M\)](#)


Protein sequence coverage: 12%

Matched peptides shown in **bold red**.


```

1  GPTFLPIQGH ILQINWYDIC QSFTHLSKPY GPVYTLYLGR KPTVYLVHGYE
51  AVYBALVWNG KKFAQGRSLP VFDKATHWNG IIFKQKQWK WTRFSLTL
101  RNLQNGVRSI EDVQEEARQ LVEELNPGQ DKNNITVIS AA_QRVQGP
151  REX_PHALSR _ERKV_EK_L LKALLNRKTD MCRARPQTHG AVTIPHQWFT
201  ELQFENLQSF KR

```

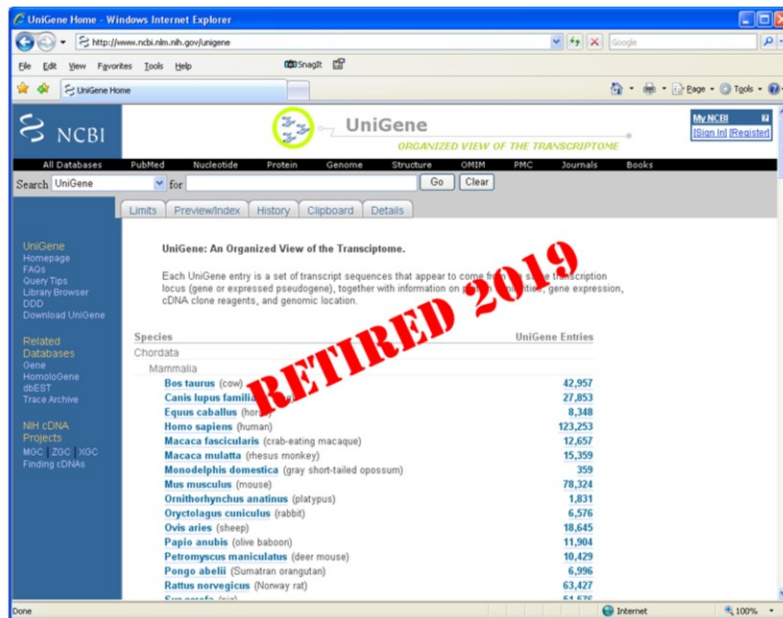

: Sequence Databases

© 2007-2022 Matrix Science



We get a protein view. This is similar to the protein view for a protein database entry, except we have drop down list for the different translation frames. For this particular entry, most of the matches have been found in reading frame 3. But, as so often happens, there is a frame shift in this entry, and there is an additional match in frame 2.

11

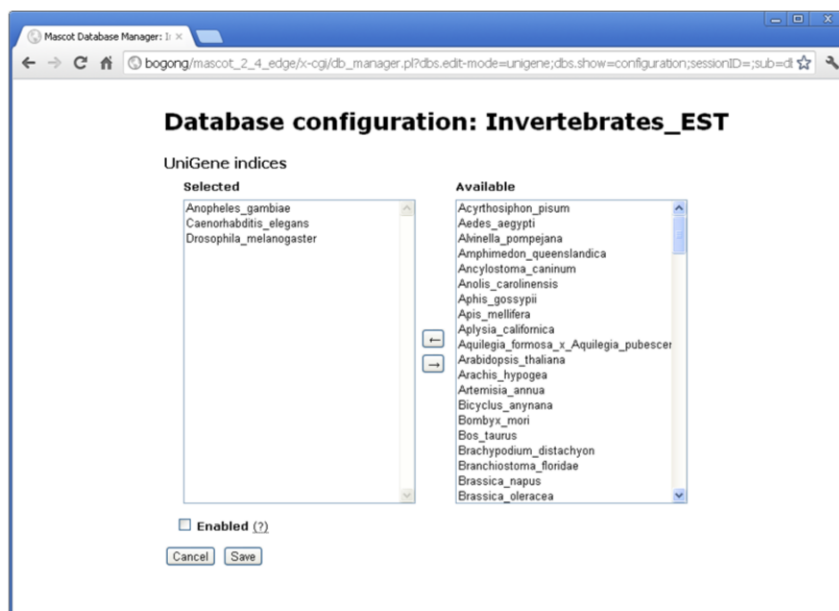


MASCOT : Sequence Databases

© 2007-2022 Matrix Science



UniGene is not a sequence database, it is an index which is created by BLASTing GenBank sequences against themselves to cluster them into gene families. NCBI retired Unigene indexes in 2019. Mascot Server 2.7 still supports the feature and the indexes are still available for download.



Unigene index files can be downloaded manually from the NCBI FTP site, but if you are using Mascot 2.4 or later, Unigene is predefined for the EST databases from both NCBI and EMBL. If enabled, index files will be downloaded automatically whenever the Fasta file is updated.

If using Mascot 2.3 or earlier, you have to make configuration changes in the database update script and mascot.dat. Details can be found in Chapter 6 of the manual and in the Mascot help page for NCBI EST

Proteins (420)
Report Builder
Unassigned (21884)
LocusMap

Protein families 1-20 (out of 346)
20 per page
1 2 3 4 5 - 18
Next
Expand all
Collapse all
Find
Clear

Accession
contains

Protein Family Summary

Format Significance threshold p< 0.05 Max. number of families AUTO [help]

Display non-sig. matches Dendrograms cut at 0

Show Percolator scores UniGene index: Mus_musculus

Preferred taxonomy All entries

Sensitivity and FDR (reversed protein sequences)

1 B2228869 347 B2228869.1 B2228869.1 NCL_GSM4_U9 Mus musculus cDNA clone (IMAGE:5152290 F...

2 B2227647 278 B2227647.1 B2227647.1 NCL_GSM4_U9 Mus musculus cDNA clone (IMAGE:5152290 F...

3 CK624204 137 CK624204.1 m18406.1 Mouse RFE/thymid. unamplified; m18406 Mus musculus cDNA d...

1 A0413050 341 A0413050.1 uc21404.1 Eugene mouse liver m184 Mus musculus cDNA clone (IMAGE:29...

2 A4189729 68 A4189729.1 m18406.1 Eugene mouse liver m184 Mus musculus cDNA clone 1...

1 CF169338 339 CF169338.1 B0812006-9 KIA Mouse Hardman Kidney cDNA Library (Lang 1) Mus muscu...

4 A0836073 83 A0836073.1 Mus musculus brain cDNA, clone WRC-7158 - 7' end...

5 B0844765 237 B0844765.1 B027982101 NCL_GSM4_Thymid Mus musculus cDNA clone (IMAGE:4930719...

3 DV057345 296 DV057345.1 MONT414_03_808.1 F1 MONT414 Mus musculus cDNA clone MONT414_03...

1 CK232350 335 CK232350.1 B0840348 Mus Musculus hematopoietic BM-HPC3 cDNA library Mus muscu...

2 B0891956 164 B0891956.1 A0890407_819280 KIA_MSC_129 Mus musculus cDNA clone (IMAGE:631...

3 B1307099 105 B1307099.1 B018758271 NCL_GSM4_Thymid Mus musculus cDNA clone (IMAGE:3482881...

1 B1145775 275 B1145775.1 B089092891 NCL_GSM4_U9 Mus musculus cDNA clone (IMAGE:5050497 F...

2 BF504121 105 BF504121.1 B0229772151 NCL_GSM4_G24 Mus musculus cDNA clone (IMAGE:4217872 S...

MASCOT : Sequence Databases © 2007-2022 Matrix Science

MATRIX SCIENCE

When Unigene is configured, we can select Mus_musculus from the drop-down list in the format controls

Proteins (905)
Report Builder
Unassigned (11844)
[permalink](#)

Protein families 1-20 (out of 895)
20 per page
1 2 3 4 5 6 ... 45 Next
Expand all Collapse all

Accession contains Find Clear

1	Mm.31018	738	Cytb5 Cytochrome b-5
2	1 Mm.330160 2 F0710191	654	Hspa3 Heat shock protein 3
		83	F0710191.1 Mus musculus mRNA 5-prime sequence 30300004660929.
3	Mm.14796	624	Mgat1 Microsomal glutathione S-transferase 1
4	Mm.15537	534	Cyp1a2 Cytochrome P450, family 1, subfamily a, polypeptide 2
5	Mm.473847	498	Transcribed locus, strongly similar to NP_044330.2 Mgat1 gene product [Mus musculus]
6	Mm.20764	481	Cyp2d9 Cytochrome P450, family 2, subfamily c, polypeptide 29
7	CB321249	477	CB321249.1 AGENCOURT_12238239 NM_MSC_136 Mus musculus cDNA clone IMAGE30...
8	Mm.289810	477	Rpl14 Ribosomal protein L14
9	Mm.425436	477	Transcribed locus, strongly similar to NP_080230.1 60S ribosomal protein L14 [Mus mus...
10	Mm.16660	434	P4hb Poly(4-hydroxylase, beta polypeptide
11	Mm.6696	411	Rdh7 Retinol dehydrogenase 7
12	1 Mm.398371 2 F0728659	407	Rpl7a Ribosomal protein L7A
		125	F0728659.1 Mus musculus mRNA 5-prime sequence from clone LADAA121YR14 (LADA...
13	Mm.328601	352	Transcribed locus, strongly similar to NP_038749.1 Rpl7a gene product [Mus musculus]
14	Mm.432030	352	Transcribed locus, strongly similar to NP_038749.1 Rpl7a gene product [Mus musculus]
15	Mm.332844	344	Cyp2a11 Cytochrome P450, family 2, subfamily a, polypeptide 11
16	Mm.20770	319	Cyp2a12 Cytochrome P450, family 2, subfamily a, polypeptide 12
17	Mm.292803	313	Ces1d Carboxylesterase 1D
18	Mm.29110	311	Ces1f Carboxylesterase 1F
19	Mm.295534	310	Ces3e Carboxylesterase 3A
20	Mm.26719	293	Hsd17b6 Hydroxysteroid (17-beta) dehydrogenase 6

20 per page
1 2 3 4 5 6 ... 45 Next
Expand all Collapse all

MASCOT

: Sequence Databases

© 2007-2022 Matrix Science



Now, using the UniGene index as a lookup table, we can transform the results of an EST search.

This is now a much clearer picture, very similar to the protein database result. Please remember that we are not clustering the database sequences into consensus sequences prior to searching. This could lead to matches being missed. UniGene is being used after the search, to map one set of accessions to a more useful set.

Threshold (E): 0.0 Cat

Query	Score	Mass	Matches	Sequences	emPAI	F
5.1 AW012478	383	15561	23 (23)	4 (4)	3.58	2
53 sameests of AW012478						
5.2 A0120761	201	25179	14 (14)	5 (5)	1.56	1
2 sameests of A0120761						
5.3 AA238951	171	23932	6 (6)	2 (2)	0.49	1
30 sameests of AA238951						
5.4 A047293	140	33937	17 (17)	5 (5)	1.02	1
2 sameests of A047293						
5.5 A132230	122	31131	9 (9)	3 (3)	0.59	6
2 sameests of A132230						
5.6 A082179	64	25761	2 (2)	2 (2)	0.45	2
2 sameests of A082179						
5.7 A029694	63	20243	6 (6)	2 (2)	0.60	2
2 sameests of A029694						

Redisplay All None

56 peptide matches (21 non-duplicate, 35 duplicate)

Auto-fit to window

Query Digest	Observed	Mr (aa)	Mr (aa)	Delta M	Score	Expect	Rank	U	1	2	3	4	5	6	7	Peptide
d1647	521.2416	1560.7029	1559.8187	0.8842	59	0.00045	1	U								K.IKISQSTFWPFE.A
d1703	525.4368	1573.3479	1572.7404	0.5924	71	0.00024	1	U								K.KALVHWRSEFPAH.G
d1541	540.2047	1678.4249	1678.3285	0.0964	54	0.00091	1	U								R.LCGRDLAR.W
d1603	541.3461	1680.7177	1680.4059	0.3118	44	0.007	1	U								K.YVDFYAR.V
d1790	577.8097	1553.8449	1553.4045	0.4404	49	0.032	1	U								R.GFFPAHAE.I
d1581	586.0058	1549.3970	1549.5994	0.1976	30	0.028	1	U								R.GFFPAHAE.I + Oxidation (H)
d12121	739.5240	1477.0534	1476.4634	0.5900	60	7.2e-05	1	U								K.OTTVITSLAVFLR.D
d15135	739.6176	1477.2007	1476.8108	0.3868	60	0.0024	1	U								R.DFDITVLIR.Q
d13443	745.3500	1489.8854	1489.8586	0.0268	73	0.00021	1	U								K.OTTVITSLAVFLR.D
d13445	810.6440	1528.8257	1528.7756	0.5000	59	0.0019	1	U								K.KALVHWRSEFPAH.G
d13473	845.5187	1529.6228	1529.7764	0.1472	90	3.3e-04	1	U								K.KALVHWRSEFPAH.G
d13952	773.1436	1544.2725	1543.8501	0.4225	61	0.00023	1	U								VYDFYITVLIR.K
d1247	781.1420	1540.2488	1539.8187	0.4311	90	5.3e-04	1	U								K.IKISQSTFWPFE.A
d12484	787.0444	1573.3743	1572.7404	0.6040	43	0.041	1	U								K.KALVHWRSEFPAH.G
d11514	804.1043	1606.1980	1605.8232	0.3747	45	0.014	1	U								K.KALVHWRSEFPAH.G
d12332	854.8924	1461.0254	1460.7014	0.3910	48	0.014	1	U								K.KALVHWRSEFPAH.G
d12307	868.6096	1705.0427	1704.8402	0.2095	45	0.0030	1	U								R.VGRAGLVEELR.A
d13549	614.1400	1639.4585	1638.8402	0.6183	42	0.038	1	U								K.KALVHWRSEFPAH.G
d14750	645.2435	1532.7486	1532.0732	0.6754	50	0.0017	1	U								K.OTTVITSLAVFLR.D
d15244	776.1400	1533.4350	1532.7404	0.6936	43	0.023	1	U								R.FIDLITSLAVFLR.D

36 subunits and intersections (215 subunit proteins in total)

The protein family summary groups entries together, but it can only connect overlapping entries which have at least one shared peptide match, so it will sometimes fail.

There are seven proteins entries grouped together in protein family 5 from the EST search. The entry names give no clue as to the protein function.

Proteins (905)
Report Builder
Unassigned (31844)
[Permalink](#)

Protein families 1-20 (out of 895)
20 per page
1 2 3 4 5 6 ... 45 Next
Expand all Collapse all

Accession contains Find Clear

1 Mm.31018 738 Cyb5 Cytochrome b-5

2 1 Mm.330160 654 Hspa5 Heat shock protein 5
2 FO710191 83 FO710191:1 Mus musculus mRNA 5-prime sequence 3030000046609839.

3 Mm.14796 624 Mgst1 Mitochondrial glutathione S-transferase 1

4 Mm.15537 534 Cyp1a2 Cytochrome P450, family 1, subfamily a, polypeptide 2

5 Mm.473847 498 Transcribed locus, strongly similar to NP_064330.2 Mgst1 gene product [Mus musculus]

6 Mm.20764 481 Cyp2c29 Cytochrome P450, family 2, subfamily c, polypeptide 29

	Score	Mass	Matches	Sequences	emPAI	F
6.1 Mm.20764	481		0	40 (40)	9 (9)	Cyp2c29 Cytochrome P450, family 2, subfamily c, polypeptide 29

40 peptide matches (12 non-duplicate, 28 duplicate)
Auto-fit to window

Query	Dupes	Observed	Hr (expt)	Hr (calc)	Delta	M	Score	Expect	Rank	U	Peptide
4447	2	521.2416	1560.7029	1559.8187	0.8842	0	59	0.00065	1	U	K.NISQSFTNFSK.A
5544	4	540.3247	1078.6349	1078.5385	0.0964	0	54	0.0091	1	U	R.ICAGEGLAR.M
5603	2	541.3661	1080.7177	1080.6059	0.1118	0	44	0.027	1	U	K.YPDVTAK.V
7790		577.9297	1153.8449	1153.6045	0.2404	0	49	0.032	1	U	R.GSFFPAEK.M
8340		586.0058	1169.9970	1169.5994	0.3976	0	33	0.028	1	U	R.GSFFPAEK.M + Oxidation (D)
81338		739.6176	1477.2207	1476.8108	0.4100	0	63	0.0026	1	U	R.DFIDYLLIK.Q
20247	7	781.1422	1560.2698	1559.8187	0.4511	0	90	5.3e-06	1	U	K.NISQSFTNFSK.A
21390	6	536.3278	1605.9617	1605.8232	0.1384	1	46	0.016	1	U	K.EALIDRGEFAGR.G
21414	1	804.1063	1606.1980	1605.8232	0.3747	1	45	0.016	1	U	K.EALIDRGEFAGR.G
23907		868.6286	1735.2427	1734.8402	0.4025	0	61	0.0032	1	U	R.VQREAGCLVEELR.R
25549		614.1602	1839.4589	1838.9602	0.4986	1	42	0.038	1	U	K.RSDYHFFSTOK.R
26750	6	645.2635	1932.7686	1932.0772	0.6914	0	50	0.0017	1	U	K.GTTVTITSLSSVLDSK.E

1 subset or intersection (1 subset protein in total)

MASCOT : Sequence Databases
© 2007-2022 Matrix Science
MATRIX SCIENCE

However, when we look at the UniGene report, we find that many of these matches all belong to the same gene, for Cytochrome P450.

In this case there was some over grouping of proteins with shared peptides, and these have been split off into separate protein families.

The other advantage of Unigene is that it gives us the more useful descriptions.

Mouse Genome Statistics

- **2.7×10^9 bases**

(Mus_EST is $\sim 2.2 \times 10^9$ bases)

- **5.4×10^9 residues in 6 frame translation**

- **99.75% of translated sequence is non-coding**

- **$\sim 1.5 \times 10^5$ tryptic limit peptides of 1500 Da \pm 0.5**

- **$\sim 6 \times 10^7$ no-enzyme peptides of 1500 Da \pm 0.5**

MASCOT

: *Sequence Databases*

© 2007-2022 Matrix Science

 **MATRIX
SCIENCE**

We can also perform MS/MS searches on the raw genomic sequence data. Let's just look at some numbers for the assembled mouse genome.

The mouse genome assembly is approximately 2.7 billion bases, which makes it a little larger than Mus_EST.

Since we must translate in all 6 reading frames, this corresponds to 5.4 billion amino acid residues.

In the mouse genome, only 1.5% of the sequence codes for proteins. This means that 99.75% of the 6 frame translation is non-coding and simply contributes to the background of random matches. This is a good test of the discrimination of the scoring scheme.

If we are matching MS/MS data from a tryptic peptide of nominal mass 1500 Da against the mouse genome, we are going to have to test 150 thousand peptides. Which sounds bad,

but is not nearly as bad as the no-enzyme case where we have to test 60 million!

U.S. National Library of Medicine

National Center for Biotechnology Information

[Log in](#)

Genome Data Viewer

GDV supports the exploration and analysis of *NCBI-annotated* and selected non-NCBI annotated eukaryotic genome assemblies. Currently, assemblies from over 1510 organisms are available.

Switch view

Search organisms

To view more organisms in the tree, click on nodes that have "+" signs. Press and hold the "+" to expand and reveal all the subgroups.
 Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.

Mus musculus (house mouse)

Search in genome

Examples: GRCm39, chr1:113043000-113056000, DNA repair

Assembly

Assembly details

Name

GRCm39

RefSeq accession

GCF_000001635.27

GenBank accession

GCA_000001635.9

Submitter

Genome Reference Consortium

Level

Chromosome

Category

Reference genome

Replaced by

GCF_000001635.26

Annotation details

Annotation Release

109

Release date

Sep 21, 2020

: **Sequence Databases**

© 2007-2022 Matrix Science

You can download the mouse genome sequences from NCBI.

National Library of Medicine
National Center for Biotechnology Information

[Datasets homepage](#) / [Assembly](#) / [GRCm39](#)

Genome assembly GRCm39 reference

Download

Reference sequence	RefSeq GCF_000001635.27
Submitted sequence	GenBank GCA_000001635.9
Taxon	<i>Mus musculus</i> (house mouse)
Strain	C57BL/6J
Submitter	Genome Reference Consortium
Date	Jun 24, 2020

[View the legacy Assembly page](#)

Assembly statistics

These statistics describe the nuclear genome of the reference sequence, GCF_000001635.27

Genome size	2.7 Gb
Number of chromosomes	21
Number of scaffolds	101
Scaffold N50	106.1 Mb
Scaffold L50	11
Number of contigs	305
Contig N50	59.5 Mb
Contig L50	15
GC percent	41.5
Assembly level	Chromosome

Download

Download a data package for GCF_000001635.27

Select file types - estimated size 728 Mb

- ☒ Genomic sequence, (FASTA)
- ☐ Annotated features (GTF)
- ☐ Annotated features (GFF3)
- ☐ Sequence and annotation (GBFF)
- ☐ Transcripts (FASTA)
- ☐ Genomic CDS (FASTA)
- ☐ Proteins (FASTA)

Your selected data will be downloaded as a ZIP archive

Name your file

Cancel Download

MASCOT
: Sequence Databases

© 2007-2022 Matrix Science

MATRIX SCIENCE

We chose the assembled chromosomes, 24 files. Although you could search this as a 24 entry database, this is not memory efficient, so we used the script mentioned earlier to split the chromosome sequences into overlapping segments of 12 kb

MASCOT Search Results

User: [blank]
 Search title: #P622008 Gensonic Mouse
 MS data file: D:\P622008\mgf\merged.mgf
 Database: Mus_musculus_GRCm39.genes: 20220435 (1,366,368 sequences, 5,530,616,312 residues)
 Timestamp: 6 May 2022 at 23:53:57 GMT

For search: ☒ All ☐ Non-significant ☐ Unassigned ☒ ☐ Export As: **XML**

For what you expected? [View report summary](#)

Search parameters
 Peptide distribution
 Identification statistics
 Legend

Protein Family Summary

Form: Max. number of families: ☒ ☐ ☐
 Display non-ig. matches: ☐ Dendrogram cut at:
 Show Percolator scores: ☐
 Preferred taxonomy:

Sensitivity and FOM (reversed protein sequences)

Proteins (312) [Link to results](#)

Protein families 1 - 10 (out of 312)

10 per page: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312

1 CH000995.3_3089 550 [View](#) [Download](#) [Details](#) [Link to results](#)
 2 CH001010.3_3014 56 [View](#) [Download](#) [Details](#) [Link to results](#)
 3 CH000994.3_3075 95 [View](#) [Download](#) [Details](#) [Link to results](#)

1 CH000999.3_31531 551 [View](#) [Download](#) [Details](#) [Link to results](#)
 2 CH001011.3_3075 452 [View](#) [Download](#) [Details](#) [Link to results](#)
 3 CH001012.3_3090 725 [View](#) [Download](#) [Details](#) [Link to results](#)
 4 CH001010.3_3076 394 [View](#) [Download](#) [Details](#) [Link to results](#)
 5 CH001011.3_3075 325 [View](#) [Download](#) [Details](#) [Link to results](#)
 6 CH001011.3_3044 309 [View](#) [Download](#) [Details](#) [Link to results](#)
 7 CH001012.3_3026 280 [View](#) [Download](#) [Details](#) [Link to results](#)
 8 CH001011.3_3042 129 [View](#) [Download](#) [Details](#) [Link to results](#)

1 CH000998.3_31531 276 [View](#) [Download](#) [Details](#) [Link to results](#)
 2 CH001014.3_30038 275 [View](#) [Download](#) [Details](#) [Link to results](#)

10 per page: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312

For what you expected? [View report summary](#)

Mascot: [Help](#) [View the Mascot website](#)

MASCOT

: Sequence Databases

© 2007-2022 Matrix Science



This is the result of searching our data against the mouse genome assembly. If you thought the Mus_EST entry titles were uninformative, how much worse is this?

Protein families 1-10 (out of 312)

Accession: Find

Threshold (E): 0.01

Accession	Score	Mass	Matches	Sequences	seqP2 F
1 CH000995.3_2889	650	475111	29 (2)	11 (1)	0.52
2 CH001010.3_2014	56	475792	2 (2)	2 (2)	0.82
3 CH000995.3_2875	56	475792	2 (2)	2 (2)	0.82

25 peptide matches (17 non-duplicate, 18 duplicate)

Query Index	Observed	NO (exp1)	NO (exp2)	Delta W Score	Report	Rank	Y	Z	Page 1 of 4
q11010 P1	602-6702	1205-1208	1205-1208	0.0007	0	0.004	1	1	1
q11010 P2	611-6703	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P3	620-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P4	629-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P5	638-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P6	647-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P7	656-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P8	665-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P9	674-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P10	683-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P11	692-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P12	701-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P13	710-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P14	719-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P15	728-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P16	737-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P17	746-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P18	755-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P19	764-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P20	773-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P21	782-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P22	791-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P23	800-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P24	809-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1
q11010 P25	818-6702	1205-1208	1205-1208	0.0004	0	0.004	1	1	1

Pro subunits and interactions (13 subunit proteins in total)

Accession	Score	Mass	Matches	Sequences	seqP2 F
1 CH000995.3_2889	551	1001000	1001000	1001000	0.00
2 CH001010.3_2014	425	1001000	1001000	1001000	0.00
3 CH001010.3_2014	425	1001000	1001000	1001000	0.00
4 CH001010.3_2014	425	1001000	1001000	1001000	0.00
5 CH001010.3_2014	425	1001000	1001000	1001000	0.00
6 CH001010.3_2014	425	1001000	1001000	1001000	0.00
7 CH001010.3_2014	425	1001000	1001000	1001000	0.00
8 CH001010.3_2014	425	1001000	1001000	1001000	0.00
9 CH001010.3_2014	425	1001000	1001000	1001000	0.00
10 CH001010.3_2014	425	1001000	1001000	1001000	0.00
11 CH001010.3_2014	425	1001000	1001000	1001000	0.00
12 CH001010.3_2014	425	1001000	1001000	1001000	0.00
13 CH001010.3_2014	425	1001000	1001000	1001000	0.00

MASCOT

: Sequence Databases

© 2007-2022 Matrix Science



If you click on an accession number link, for a protein view report, you can get either the standard protein view report or an alternative

MASCOT : Sequence Databases © 2007-2022 Matrix Science

This is the peptide match results formatted as an EMBL / GenBank format feature table. This may not look very friendly, but the advantage of this report is that it can be read into a standard genome browser.

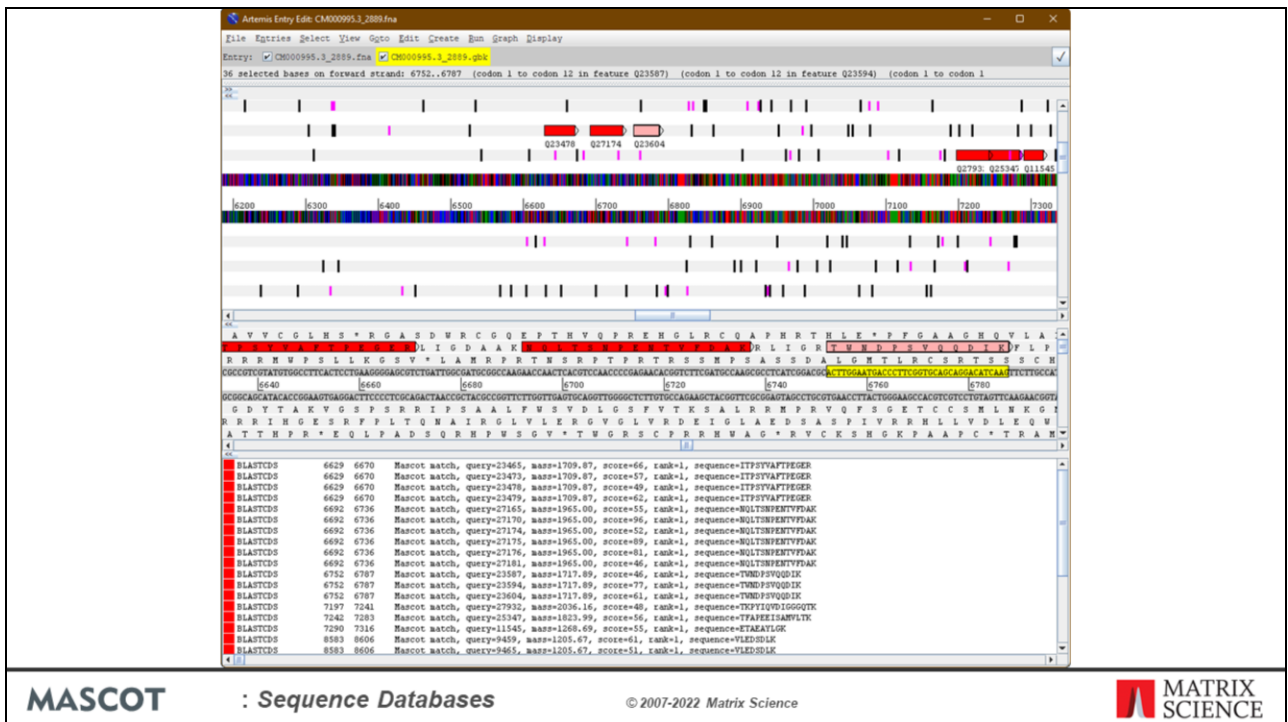
To enable this feature add the “FeatureTableLength” parameter to the options section of the mascot.dat file using the Configuration Editor->Configuration Options editor or a text editor. Set the value to less than the number of bases that the genomic was split into. A FeatureTableLength 10000 is a good value.

The screenshot displays the Artemis genome browser interface. At the top, there's a navigation bar with links: Downloads, Further information, Contact, Publications, Programmes and Facilities. Below this, a red banner contains the text "Tool Annotations Pipelines and Modules". The main heading is "Artemis", followed by a description: "Genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation." Below this, it states "Artemis is a free genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation." and "About".

The central part of the interface shows a sequence viewer with a feature table. The feature table has columns for "Name", "Feature", "Start", "End", "Score", "Gene", and "Accession". The "Name" column contains "Gene", "Feature", "Start", "End", "Score", "Gene", and "Accession". The "Feature" column contains "Gene", "Feature", "Start", "End", "Score", "Gene", and "Accession". The "Start" column contains "1", "1", "1", "1", "1", "1", "1". The "End" column contains "100", "100", "100", "100", "100", "100", "100". The "Score" column contains "100", "100", "100", "100", "100", "100", "100". The "Gene" column contains "Gene", "Feature", "Start", "End", "Score", "Gene", and "Accession". The "Accession" column contains "1", "1", "1", "1", "1", "1", "1".

At the bottom of the interface, there's a footer with the text "MASCOT : Sequence Databases", "© 2007-2022 Matrix Science", and the "MATRIX SCIENCE" logo.

Here's the result of reading the feature table containing the Mascot peptide matches into Artemis.



In the upper third, we have a low resolution view. This can be zoomed out to show an entire sequence as a single strip. We have the forward and complementary DNA strands, and the 6 frame translation. The vertical bars are start (purple) and stop (black) codons.. Individual Mascot peptide matches are shown in red or pink when selected. This particular gene has 11 peptide matches.

The middle third is a similar arrangement, but at high enough resolution to see individual bases and residues.

Finally, the lower third shows a tabular view of the feature table. When a match is selected, it is highlighted in all three views, and we can see the spectrum number, sequence, molecular weight, Mascot score, etc.

Not only does this allow us to zoom and pan around these extremely long sequences, it also allows us to view the peptide matches found by Mascot in the context of all the existing annotations. This gives us a powerful way to present the results of MS based searching complete genomes.

Mouse UniProt vs. EST vs. Genome

▼ Search parameters

Type of search	: MS/MS Ion Search
Enzyme	: Trypsin/P
Fixed modifications	: O^6 TRAQ4plex (K), O^6 TRAQ4plex (N-term), O^6 Methylthio (C)
Variable modifications	: O^6 Acetyl (Protein N-term), O^6 Gln->pyro-Glu (N-term Q), O^6 Oxidation (M)
Mass values	: Monoisotopic
Protein mass	: Unrestricted
Peptide mass tolerance	: ± 0.9 Da
Fragment mass tolerance	: ± 0.6 Da
Max missed cleavages	: 1
Instrument type	: ESI-TRAP
Number of queries	: 33,191

Database	Size in residues	Average score threshold	Number of PSMs (1% FDR)
Uniprot mouse	2.8×10^7	37	1834
EST mouse	4.5×10^9	59	675
Mouse genome	5.5×10^9	59	548

MASCOT

: Sequence Databases

© 2007-2022 Matrix Science



All well and good, but which database gives the most matches? We searched a larger dataset against all 3 databases. The data was the public iPRG2008 dataset distributed by ABRF.

There is a big drop in the number of matches between Uniprot mouse and EST mouse. The reason is mainly that EST mouse is a much bigger database, by more than a factor of 100. This means that the score thresholds are approx 22 higher, and we lose all the weaker matches, that had scores between 37 and 59. Yes, there may be additional matches in EST, not found in Uniprot, but the net change is highly negative.

You can see at a glance that the mouse genome is even worse. This is not because of a still higher threshold; although the database is slight larger than Mus_EST the thresholds are the same. One reason is that a proportion of potential matches are missed because they are split across exon-intron boundaries. Based on average peptide length, approx 20% of matches would be lost for this reason. In this particular example, the difference is just under 20% at 18.8%. The other factor is that the mouse genome is only 1.5% coding sequence, and represents a single consensus genome. EST is 100% coding sequence and represents a wide range of SNPs and variants.

neXtProt vs. EST vs. Genome

- Searching complete chromosomes is possible, but unwieldy.
- Scoring statistics for assembled genome very similar to Mus_EST, but
 - the genome is a single consensus sequence, Mus_EST represents many variants
 - Mus_EST is 100% coding, MG assembly is 1.5% coding
 - lose approx 20% of matches because they straddle an exon - intron boundary

• In general, Mus_EST is a better choice

• References

Choudhary, J. S., Blackstock, W. P., Creasy, D. M. and Cottrell, J. S. (2001). Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics*, 1, 651-667.
Choudhary, J. S., Blackstock, W. P., Creasy, D. M. and Cottrell, J. S. (2001). Matching peptide mass spectra to EST and genomic DNA databases. *Trends in Biotechnology*, 19, S17-S22.

So, these are our conclusions for the mouse genome, and the same considerations probably hold for other large mammalian genomes.

Plant and bacterial genomes are a different matter. If the species is not well represented in the protein databases, there is a much stronger need to search EST or genomic databases