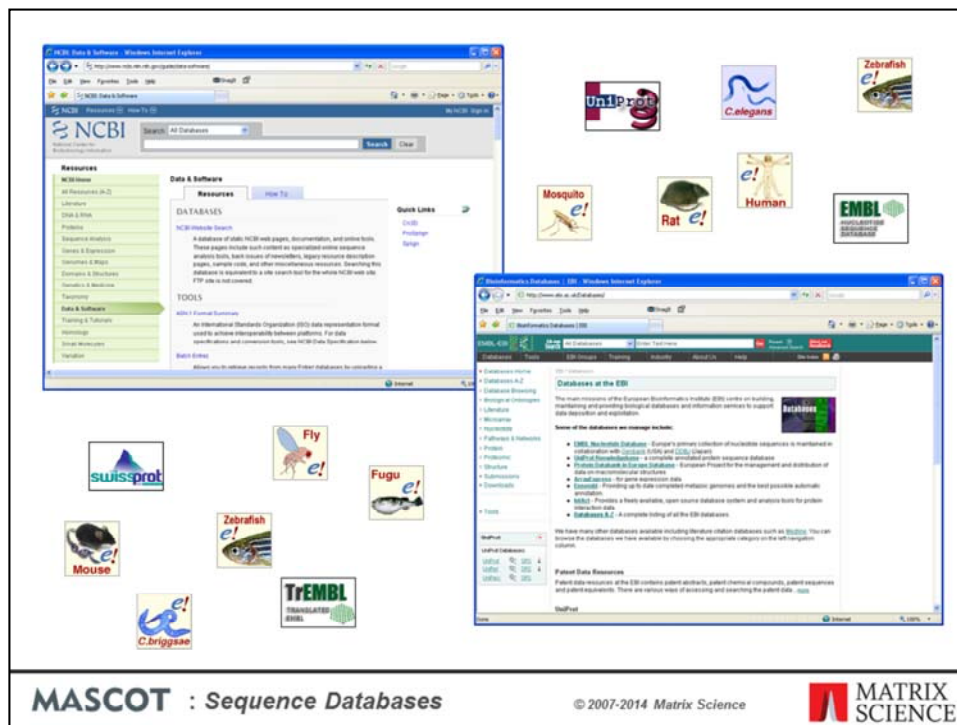


Sequence Databases

MASCOT

 MATRIX
SCIENCE



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 (limit of 64 in Mascot 2.2 and earlier).

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 (~550,000 entries)

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

NCBI nr, UniRef100 (~48,000,000 entries)

- Comprehensive, non-identical

UniRef90, UniRef50, etc.

- UniRef100 better for MS/MS; need explicit sequences

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

MASCOT : *Sequence Databases*

© 2007-2014 Matrix Science



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

Swiss-Prot 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.

The comprehensive, non-identical databases are a good choice for MS/MS searching if you don't want to miss any matches. NCBI nr and UniRef100 both aim to include explicit representations of all known protein sequences.

If you search a non-redundant version, such as UniRef90, 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 2014)

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
 A S P S D # Q Q I S W Q K L Y C P E E S G G Q Y S T I
 Q A H Q T N S R S L G R N S T S O K R V G A N I Q H
 CAAGCCCATCAGACTAACAGCAGATCTTTGGCAGAACTCTACAAGCCGAAGAGAGTGGGGGCAATATTCAACATT
 [299200] [299210] [299220] [299230] [299240] [299250] [299260] [299270]
 TTTCGGGTAGTCTGATTGCTGCTCAGAGAACCGTCTTTGAGATGTTCTCTCTCACCCCGGTATTAAGTTGTAT
 A W * V L L L L D R P L F E V L W F L T P A L I * C H
 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: TTTTNTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCAAAAAAGGGGGGGGGGGGGGGGGGG
Base 2: TTTTCCCAAAAGGGGTTTTCCC AAAAGGGGTTTTCCC AAAAGGGGTTTTCCC AAAAGGGG
Base 3: TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

* = stop

MASCOT : *Sequence Databases*

© 2007-2014 Matrix Science



MATRIX
SCIENCE

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

© 2007-2014 Matrix Science



All the genomes in GenBank are translated into protein sequences in NCBItr. 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

MASCOT : *Sequence Databases*

© 2007-2014 Matrix Science



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 : ms_user
 Email : ms@localhost
 Search title : BART - 8
 Database : uniprot_human_human_20140709 (89005 sequences; 35230190 residues)
 Timestamp : 9 Sep 2014 at 15:15:47 GMT
 Protein hits :

Accession	Description
P05187	Alkaline phosphatase, placental type OS=Homo sapiens GN=ALPP PE=1 SV=2
P10676	Alkaline phosphatase, placental-like OS=Homo sapiens GN=ALPL1 PE=1 SV=4
J3Q8D1	Tyrosine-protein kinase ves OS=Homo sapiens GN=VRS1 PE=1 SV=1
P08168	Keratin, type II cytoskeletal 1 OS=Homo sapiens GN=KRT1 PE=3 SV=6
A1L4C2	Brain-specific angiogenesis inhibitor 1-associated protein 2 OS=Homo sapiens GN=BAIAP2 PE=1 SV=1
P15588	Keratin, type II cytoskeletal 2 epidermal OS=Homo sapiens GN=KRT2 PE=1 SV=2
D08488	Oligo(2)-diphospholipidase--protein glycosyltransferase subunit 2 OS=Homo sapiens GN=RPNG2 PE=1 SV=3
P15527	Keratin, type I cytoskeletal 9 OS=Homo sapiens GN=KRT9 PE=1 SV=3
P12513	Proto-oncogene tyrosine-protein kinase Src OS=Homo sapiens GN=SRC PE=1 SV=3
B18135	Pyruvate kinase (fragment) OS=Homo sapiens GN=PKM PE=1 SV=1
P16188	Carboxypeptidase Y OS=Homo sapiens GN=CPY PE=1 SV=2
Q50898	Brain-specific angiogenesis inhibitor 1-associated protein 2-like protein 1 OS=Homo sapiens GN=BAIAP2L1 PE=1 SV=2
B07248	Tyrosine-protein kinase Lyn OS=Homo sapiens GN=LYN PE=1 SV=3
P09523	Intestinal-type alkaline phosphatase OS=Homo sapiens GN=ALPI PE=1 SV=2
P13166	Solute carrier family 2, facilitated glucose transporter member 1 OS=Homo sapiens GN=SLC2A1 PE=1 SV=2
Q12728	Keratin, type II cytoskeletal 1b OS=Homo sapiens GN=KRT77 PE=2 SV=3
B07253	Tyrosine-protein kinase HCK OS=Homo sapiens GN=HCK PE=1 SV=1
P10750	Cyclin-dependent kinase 9 OS=Homo sapiens GN=CDK9 PE=1 SV=3
Q14004	Cyclin-dependent kinase 13 OS=Homo sapiens GN=CDK13 PE=1 SV=2
J2Q607	Cyclin-dependent kinase 12 (fragment) OS=Homo sapiens GN=CDK12 PE=1 SV=1
E19022	Keratin, type II cytoskeletal 5 (fragment) OS=Homo sapiens GN=KRT5 PE=1 SV=1
Q9NE75	Retinoic acid-induced protein 3 OS=Homo sapiens GN=RP3A PE=1 SV=2
A48959	Proline and serine-rich protein 2 (fragment) OS=Homo sapiens GN=PROSER2 PE=4 SV=1
P08159	Keratin, type II cytoskeletal 6B OS=Homo sapiens GN=KRT6B PE=1 SV=5
E19883	Keratin, type II cytoskeletal 4 OS=Homo sapiens GN=KRT4 PE=1 SV=1
Q01546	Keratin, type II cytoskeletal 2 oral OS=Homo sapiens GN=KRT76 PE=1 SV=2
C13250	Uncharacterized protein OS=Homo sapiens GN=STAP4 PE=4 SV=1
G19156	Polypeptide N-acetylglucosaminyltransferase 2 soluble form OS=Homo sapiens GN=GALNT2 PE=1 SV=1
P13059	Keratin, type I cytoskeletal 13 OS=Homo sapiens GN=KRT13 PE=1 SV=4
B07205	Uncharacterized protein (fragment) OS=Homo sapiens PE=4 SV=1
B07250	Beta-2-syntrophin (fragment) OS=Homo sapiens GN=SYNTB2 PE=1 SV=1
C13203	IM domain-containing protein 1 (fragment) OS=Homo sapiens GN=IMD1 PE=1 SV=1
Q08176	Coronin-18 OS=Homo sapiens GN=COR18 PE=1 SV=1
B07223	Protein disulfide-isomerase (fragment) OS=Homo sapiens GN=PDIB PE=1 SV=1
B08107	Ubiquitin-60S ribosomal protein L40 (fragment) OS=Homo sapiens GN=UBA52 PE=4 SV=1

MASCOT : Sequence Databases © 2007-2014 Matrix Science **MATRIX SCIENCE**

To illustrate the features of the different types of database, we first searched a very small dataset of a few hundred MS/MS spectra against a protein database, the Uniprot complete human proteome. There are significant matches to some 36 human proteins

Mascot Search Results

User : ms_user
 Email : ms@localhost
 Search title : BART - 8
 Database : EST human human_20140903 (52229330 sequences; 8864075794 residues)
 Timestamp :
 Protein hits :

Accession	Description
gi149053565	BX458398 Homo sapiens PLACENTA Homo sapiens cDNA clone CS00E002V021 5'-PRIME, mRNA sequence
gi149053561	602631560F1 HCT_C8AP_5h3 Homo sapiens cDNA clone IPAGE:4778638 5', mRNA sequence
gi149053560	AL555511 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS00I076V024 5'-PRIME, mRNA sequence
gi149053557	002631554 UTERUS Homo sapiens cDNA clone UTERUS022444 5', mRNA sequence
gi149053557	602627035F1 HCT_C8AP_5h4 Homo sapiens cDNA clone IPAGE:4751984 5', mRNA sequence
gi149053557	601434343F1 HCT_HDC_72 Homo sapiens cDNA clone IPAGE:3932498 5', mRNA sequence
gi149053555	DA833050 PLAC1 Homo sapiens cDNA clone PLAC1010376 5', mRNA sequence
gi149053555	CR988750 RZPD no.9017 Homo sapiens cDNA clone RZPD09017X062 5', mRNA sequence
gi149053555	C10-000330-20101-050-F02 C000330 Homo sapiens cDNA, mRNA sequence
gi149053555	DB462453 RIKEN full-length enriched human cDNA library, testis Homo sapiens cDNA clone HD1309H04 5', mRNA sequence
gi149053555	AL578382 Homo sapiens HELA CELLS COT 25-NORMALIZED Homo sapiens cDNA clone CS00K010Y004 3'-PRIME, mRNA sequence
gi149053555	602597767F1 HCT_HDC_79 Homo sapiens cDNA clone IPAGE:4604021 5', mRNA sequence
gi149053555	602651220F1 HCT_HDC_47 Homo sapiens cDNA clone IPAGE:4761943 5', mRNA sequence
gi149053555	DA833815 PLAC1 Homo sapiens cDNA clone PLAC1009867 5', mRNA sequence
gi149053555	60304479F1 HCT_HDC_115 Homo sapiens cDNA clone IPAGE:5175730 5', mRNA sequence
gi149053555	603647323F1 HCT_HDC_38 Homo sapiens cDNA clone IPAGE:5439164 5', mRNA sequence
gi149053555	ADENOCOURT_6626561 HCT_HDC_118 Homo sapiens cDNA clone IPAGE:5735492 5', mRNA sequence
gi149053555	c-EST0096024 522501601 Homo sapiens cDNA clone 522501601-96-C08 5', mRNA sequence
gi149053555	DA007848 FHS322 Homo sapiens cDNA clone FHS322005139 5', mRNA sequence
gi149053555	DA022749 ASTRO2 Homo sapiens cDNA clone ASTRO2004865 5', mRNA sequence
gi149053555	602633634F1 HCT_C8AP_5h3 Homo sapiens cDNA clone IPAGE:4778810 5', mRNA sequence
gi149053555	NCI-H9372-241329-111 NCI-H9372 Homo sapiens cDNA, mRNA sequence
gi149053555	DA567559 HELAC2 Homo sapiens cDNA clone HELAC2000021 5', mRNA sequence
gi149053555	c-EST015315 15MLK1 Homo sapiens cDNA clone 15MLK1-20-E11 5', mRNA sequence
gi149053555	602632960F1 HCT_C8AP_5h3 Homo sapiens cDNA clone IPAGE:4778043 5', mRNA sequence
gi149053555	DB277466 UTERUS Homo sapiens cDNA clone UTERUS003128 5', mRNA sequence
gi149053555	HP267404 RIKEN full-length enriched human cDNA library, thymus Homo sapiens cDNA clone H050093110, mRNA sequence
gi149053555	DA383601 BRTN2 Homo sapiens cDNA clone BRTN2017967 5', mRNA sequence
gi149053555	C0454472 Homo sapiens ORESTES from keratinocytes Homo sapiens cDNA, mRNA sequence
gi149053555	V43 Human epidermis granular keratinocytes Homo sapiens cDNA, mRNA sequence
gi149053555	2445F04.r1 Saarens pregnant uterus fibroblast Homo sapiens cDNA clone IPAGE:485791 5' similar to gh:Y0N282 RIBOPHOSPHIN II PRECURSOR (HPIPH2), mRNA se
gi149053555	C0452446 Homo sapiens ORESTES from keratinocytes Homo sapiens cDNA, mRNA sequence
gi149053555	602626883F1 HCT_C8AP_5h4 Homo sapiens cDNA clone IPAGE:4751637 5', mRNA sequence
gi149053555	ADENOCOURT_7941061 HCT_HDC_112 Homo sapiens cDNA clone IPAGE:6110481 5', mRNA sequence
gi149053555	BP216200 Sugano cDNA library, corpus callosum Homo sapiens cDNA clone CC060223 5', mRNA sequence
gi149053555	ADENOCOURT_6640963 HCT_HDC_39 Homo sapiens cDNA clone IPAGE:5434174 5', mRNA sequence
gi149053555	ADENOCOURT_6640963 HCT_HDC_39 Homo sapiens cDNA clone IPAGE:5434174 5', mRNA sequence

MASCOT : Sequence Databases © 2007-2014 Matrix Science **MATRIX SCIENCE**

With EST_human, 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.

Peptide Summary Report

54.243.190.62/mascot/cgi/master_results.pl?file=.%2fdata%2f20140905%2fH001270.dat_ignorescorebelow=0_minpeplen=7_prefertaxonomy=&=&

Select All Select None Search Selected Error tolerant Archive Report

1. [gi147031565](#) Mass: 34632 Score: 419 Matches: 8(5) Sequences: 7(4) mpAI: 0.62 Frame: 2
BX458398 Homo sapiens PLACENTA Homo sapiens cDNA clone CS80E002Y021 5-PRIME, mRNA sequence
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta Miss Score	Expect	Rank	Unique	Peptide
22	462.6807	923.3468	923.5116	-0.1649	0	33	35	R.FPPVALSK.T
65	567.6567	1133.2987	1133.5499	-0.2511	0	45	4.4	R.GNEVTSVQNR.A + Oxidation (M)
88	614.2001	1226.3856	1226.6329	-0.2473	0	28	1.9e+02	U K.LGPFLPLAGDR.F + Oxidation (M)
120	653.2191	1304.4057	1304.6817	-0.2760	0	87	0.00023	R.QWPTDLSASAR.F
126	726.1806	1450.3465	1450.6477	-0.3011	0	69	0.015	R.NQYSDQNPASAR.Q
208	975.8100	1949.6055	1950.0245	-0.4190	0	86	0.00019	R.NLIFLGDGQSVTYAAR.I + Oxidation (M)
210	1001.2027	2000.3508	2000.8058	-0.4150	0	65	0.022	R.GCTPDPEYPDYSGGTR.L + Oxidation (M)
212	667.8046	2000.3519	2000.8058	-0.4139	0	73	0.004	R.GCTPDPEYPDYSGGTR.L + Oxidation (M)

Proteins matching the same set of peptides:

[gi146572885](#) Mass: 35862 Score: 417 Matches: 8(5) Sequences: 7(4)
BX379970 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS80D042Y009 5-PRIME, mRNA sequence

[gi145837268](#) Mass: 35339 Score: 415 Matches: 8(5) Sequences: 7(4)
AL552555 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS80D067Y024 5-PRIME, mRNA sequence

2. [gi148051261](#) Mass: 31277 Score: 388 Matches: 7(3) Sequences: 7(3) mpAI: 0.49
602631560P1 HCT_05AP_Skin3 Homo sapiens cDNA clone IMAGE:4776638 5', mRNA sequence
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta Miss Score	Expect	Rank	Unique	Peptide
16	487.1896	972.3246	972.5240	-0.1994	0	51	0.88	R.IEISELNR.V
20	590.1804	1178.3462	1178.5931	-0.2469	0	35	37	R.YEELQITAGR.H
22	651.7282	1301.4419	1301.7076	-0.2659	0	63	0.053	R.SLDLDSIIAEVR.A
111	679.2119	1356.4092	1356.6885	-0.2792	0	75	0.0033	R.LNMLEDALQQAR.A
112	665.1406	1392.4001	1392.7249	-0.3248	1	49	1.3	U R.TNAENEFVTIDK.D
128	738.2301	1474.4457	1474.7416	-0.2959	0	53	0.56	R.HELLQQVDTSTR.T
122	738.2697	1474.5249	1474.7780	-0.2531	0	67	0.018	R.FLEQQQLQTR.H

MASCOT : Sequence Databases © 2007-2014 Matrix Science **MATRIX SCIENCE**

The master results report from the EST search looks pretty similar to the IPI 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: gi|14051361

602631560F1 NCI_CGAP_Skn3 Homo sapiens cDNA clone
IMAGE:4776638 5', mRNA sequence

Database: EST_human
Score: 386
Nominal mass (M_r): 31277
Calculated pI: 5.17
Frame: 1
Taxonomy: [Homo sapiens](#)

NB: Matches were also found in other frames indicating a possible frame shift. Only matches in frame 1 are shown in this report.

Show frame: 1 ▼

Sequence similarity is available as an [NCBI BLAST search of gi|14051361 against nr](#).

Search parameters

Enzyme: Trypsin/P: cuts C-term side of KR.
Fixed modifications: [Carbamidomethyl \(C\)](#)
Variable modifications: [Oxidation \(M\)](#)

Protein sequence coverage: 24%

Matched peptides shown in **bold red**.

1 VFFLEQQNQV LQTHMELLQQ VDTSTKML EPTFESTINI LARRVDQLK
51 DQSLGSEIK IQQMDVEDR NYVESEINR TWAKNEPVTI KKDIDGAYNG
101 NYVIGAKLIN LQGDIDFLA LVQALSDND TOTREIVIL SHGDRALDL
151 DKLIAENWQA YEDIAQKFA EASELVQSY KSLQITAKR GDSVNSFIE
201 ISKLWNVQR LRSIEDVYR QISHCQSSIS DAEGRGKAS RQFRTS_MTV
251 RIPCRRFRQ LGRLLP_LP

MASCOT : Sequence Databases

© 2007-2014 Matrix Science

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 1.

But, as so often happens, there is a frame shift in this entry, and there are additional matches in frame 3.

Mascot Search Results

User: ms_user
 Email: ms@localhost
 Search title: BART - 8
 Database: ESI human human_20140903 (52229330 sequences; 8864075794 residues)
 Timestamp:
 Protein hits:

gi 47053565	84458398	Homo sapiens PLACENTA Homo sapiens cDNA clone CS000002V021 5'-PRIME, mRNA sequence
gi 18951361	602631560F1	HCT_C8AP_5h3 Homo sapiens cDNA clone IPAGE:4776638 5', mRNA sequence
gi 45858299	AL555511	Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS001076V024 5'-PRIME, mRNA sequence
gi 183080712	00263154	UTERUS Homo sapiens cDNA clone UTERUS022441 5', mRNA sequence
gi 13310879	602627035F1	HCT_C8AP_5h4 Homo sapiens cDNA clone IPAGE:4751984 5', mRNA sequence
gi 10251178	00143454F1	NDH_FDC_72 Homo sapiens cDNA clone IPAGE:3932498 5', mRNA sequence
gi 18242655	DA833050	PLACE1 Homo sapiens cDNA clone PLACE1010376 5', mRNA sequence
gi 18282635	CR988750	RZPD no.9017 Homo sapiens cDNA clone RZPD09017X062 5', mRNA sequence
gi 14612422	C10-000330-201021-050-F02	GM00330 Homo sapiens cDNA, mRNA sequence
gi 15081228	DB462453	R1KEN Full-length enriched human cDNA library, testis Homo sapiens cDNA clone HD1309H04 5', mRNA sequence
gi 46327223	AL578382	Homo sapiens HELA CELLS COT 25-NORMALIZED Homo sapiens cDNA clone CS000010V004 3'-PRIME, mRNA sequence
gi 13361700	002507767F1	NDH_FDC_79 Homo sapiens cDNA clone IPAGE:4604021 5', mRNA sequence
gi 13314115	002651220F1	NDH_FDC_47 Homo sapiens cDNA clone IPAGE:4761943 5', mRNA sequence
gi 18238080	DA833815	PLACE1 Homo sapiens cDNA clone PLACE1009867 5', mRNA sequence
gi 15031242	60304479F1	NDH_FDC_115 Homo sapiens cDNA clone IPAGE:5175730 5', mRNA sequence
gi 15513588	60364732F1	NDH_FDC_38 Homo sapiens cDNA clone IPAGE:5439164 5', mRNA sequence
gi 15327280	ADENKOUNT_6620501	NDH_FDC_118 Homo sapiens cDNA clone IPAGE:5795492 5', mRNA sequence
gi 13180918	E-EST0066024	522501661 Homo sapiens cDNA clone 522501661-96-C08 5', mRNA sequence
gi 180536324	Q4087848	DMB322 Homo sapiens cDNA clone DM322005139 5', mRNA sequence
gi 78467671	DA022749	ASTRO2 Homo sapiens cDNA clone ASTRO2004865 5', mRNA sequence
gi 14051883	602633634F1	HCT_C8AP_5h3 Homo sapiens cDNA clone IPAGE:4778810 5', mRNA sequence
gi 10808748	NCI-H9372-241359-411	481c H9372 Homo sapiens cDNA, mRNA sequence
gi 18180925	DA567559	HELAC2 Homo sapiens cDNA clone HELAC2000021 5', mRNA sequence
gi 27297454	E-EST0153115	159LX1 Homo sapiens cDNA clone 159LX1-20-E11 5', mRNA sequence
gi 14051418	002632906F1	HCT_C8AP_5h3 Homo sapiens cDNA clone IPAGE:4778043 5', mRNA sequence
gi 18261459	DB277466	UTERUS Homo sapiens cDNA clone UTERUS003128 5', mRNA sequence
gi 180283160	HP267404	R1KEN Full-length enriched human cDNA library, thymus Homo sapiens cDNA clone HD00003110, mRNA sequence
gi 181316300	DA383601	BTHM2 Homo sapiens cDNA clone BTHM2017967 5', mRNA sequence
gi 15473122	C0454472	Homo sapiens ORESTES from keratinocytes Homo sapiens cDNA, mRNA sequence
gi 14090278	V43	Human epidermis granular keratinocytes Homo sapiens cDNA, mRNA sequence
gi 15161391	2145F04.r1	Saori pregnant uterus H00PU Homo sapiens cDNA clone IPAGE:485791 5' similar to gh:Y0N282 RIZOPHOSPHIN II PRECURSOR (HPIPH2), mRNA se
gi 154730304	C0454446	Homo sapiens ORESTES from keratinocytes Homo sapiens cDNA, mRNA sequence
gi 13311608	602626883F1	HCT_C8AP_5h4 Homo sapiens cDNA clone IPAGE:4751637 5', mRNA sequence
gi 12051601	ADENKOUNT_7941001	NDH_FDC_112 Homo sapiens cDNA clone IPAGE:6110481 5', mRNA sequence
gi 15208361	BP216200	Sugano cDNA library, corpus callosum Homo sapiens cDNA clone CCR04223 5', mRNA sequence
gi 15332986	ADENKOUNT_6640963	NDH_FDC_39 Homo sapiens cDNA clone IPAGE:5434174 5', mRNA sequence
gi 15332986	ADENKOUNT_6640963	NDH_FDC_39 Homo sapiens cDNA clone IPAGE:5434174 5', mRNA sequence

Going back to the issue of the hit list and the descriptions not saying very much. There are several problems here. One is that EST databases usually have a huge amount of redundancy, which can make for very long reports. Another problem is that the sequences tend to be short, so we don't get much grouping of peptide matches into protein matches.

To address this problem, we can use the UniGene index from the National Center for Biotechnology Information to simplify the search results.

The screenshot shows the UniGene website in a Windows Internet Explorer browser. The address bar displays "http://www.ncbi.nlm.nih.gov/uniGene". The page features the NCBI logo and the UniGene logo with the tagline "ORGANIZED VIEW OF THE TRANSCRIPTOME". A search bar is present with the text "Search UniGene" and buttons for "Go" and "Clear". Below the search bar are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details".

The main content area is titled "UniGene: An Organized View of the Transcriptome." and includes a description: "Each UniGene entry is a set of transcript sequences that appear to come from the same transcription locus (gene or expressed pseudogene), together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location." Below this is a table with two columns: "Species" and "UniGene Entries".

Species	UniGene Entries
Chordata	
Mammalia	
<i>Bos taurus</i> (cow)	42,957
<i>Canis lupus familiaris</i> (dog)	27,953
<i>Equus caballus</i> (horse)	8,348
<i>Homo sapiens</i> (human)	123,253
<i>Macaca fascicularis</i> (crab-eating macaque)	12,657
<i>Macaca mulatta</i> (rhesus monkey)	15,359
<i>Monodelphis domestica</i> (gray short-tailed opossum)	359
<i>Mus musculus</i> (mouse)	78,324
<i>Ornithorhynchus anatinus</i> (platypus)	1,831
<i>Oryctolagus cuniculus</i> (rabbit)	6,576
<i>Ovis aries</i> (sheep)	18,645
<i>Papio anubis</i> (olive baboon)	11,504
<i>Petromyscus maniculatus</i> (deer mouse)	18,429
<i>Pongo abelii</i> (Sumatran orangutan)	6,596
<i>Rattus norvegicus</i> (Norway rat)	63,427

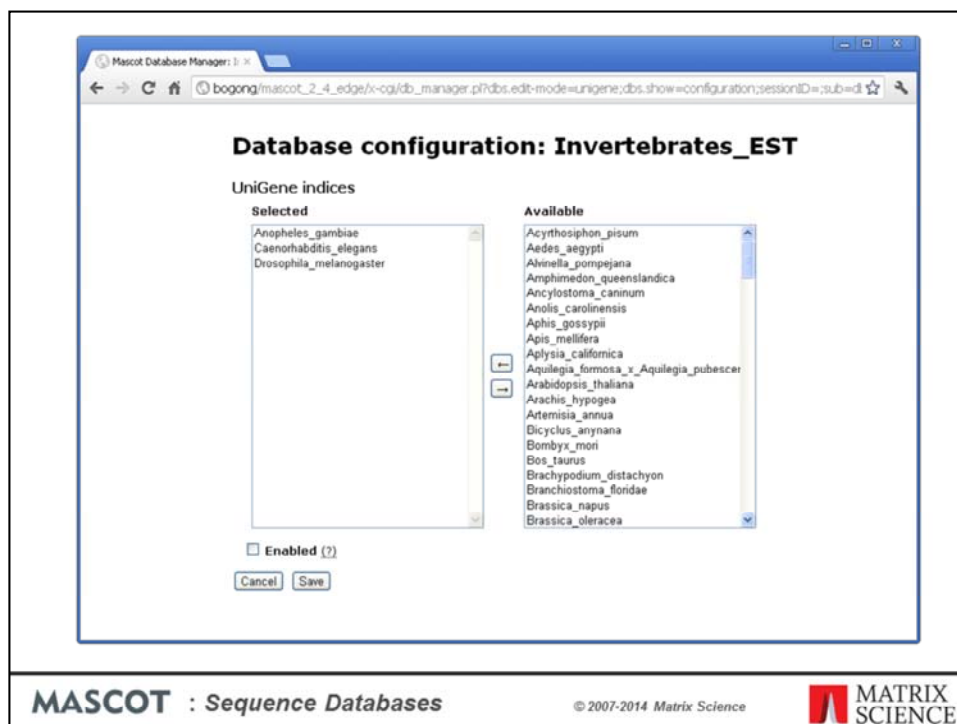
At the bottom of the browser window, the status bar shows "Done", "Internet", and "100%" zoom level.

MASCOT : Sequence Databases

© 2007-2014 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.



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

The screenshot displays the Mascot Search Results interface. On the left, a sidebar shows search parameters: User (ms_user), Email (ms@matrixscience.com), Search title (BAFT - 8), Database (EST human human_20140903 (52229330 sequences; 8864075794 residues)), Timestamp, and Protein hits. The main area shows a list of protein hits with their UniGene IDs and descriptions. On the right, a 'Format controls' panel is visible, containing a 'UniGene index' dropdown menu set to 'Homo_sapiens', a 'Max. number of' dropdown menu set to 'Homo_sapiens', an 'Ions score or expect cut-off' input field set to '0', and a 'Show sub-sets' input field set to '0'. A 'Help' link is also present.

MASCOT : Sequence Databases

© 2007-2014 Matrix Science

MATRIX SCIENCE

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

Mascot Search Results

User : ms_user
 Email : ms@localhost
 Search title : BAFT - 8
 Database : EST human human_20140903 (52229330 sequences; 8864073794 residues)
 Timestamp : 5 Sep 2014 at 15:40:15 GMT
 Protein hits :

Accession	Description
Ms.333509	ALPL2 Alkaline phosphatase, placental-like 2
Ms.284255	ALPP Alkaline phosphatase, placental
Ms.284148	YES1 Vyes-1 Yamaguchi sarcoma viral oncogene homolog 1
Ms.40823	KRT1 Keratin 1
Ms.428310	BAIAP2 BAI1-associated protein 2
Ms.416705356	B4458398 Homo sapiens PLACENTA Homo sapiens cDNA clone CS00E002YN21 5'-PRIME, mRNA sequence
Ms.307	KRT2 Keratin 2
Ms.4182422645	D4833050 PLAC1 Homo sapiens cDNA clone PLAC1010376 5', mRNA sequence
Ms.342220	PYK Pyruvate kinase, muscle
Ms.418661282	CH3-040330-260101-650-602 G40330 Homo sapiens cDNA, mRNA sequence
Ms.335650	SAC V-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)
Ms.301272	B5G B5 antigen (ON blood group)
Ms.354229	(vsi30) mRNA, 200bp
Ms.491267	LYN Vyes-1 Yamaguchi sarcoma viral related oncogene homolog
Ms.41180955	D4567559 HELAC2 Homo sapiens cDNA clone HELAC200021 5', mRNA sequence
Ms.411821169	MY267404 K12EN full-length enriched human cDNA library, thymus Homo sapiens cDNA clone H05D093110, mRNA sequence
Ms.370895	RPL2 Ribophorin II
Ms.41180617292	MY084440 K12EN full-length enriched human cDNA library, thymus Homo sapiens cDNA clone H05D112M07, mRNA sequence
Ms.320640	ACTB Actin, beta
Ms.433845	KRT5 Keratin 5
Ms.454387	CPY Carboxypeptidase Y
Ms.335132	EF1A1 Eukaryotic translation elongation factor 1 alpha 1
Ms.41385128471	MY111112 K12EN full-length enriched human cDNA library, brain Homo sapiens cDNA clone H04D023M08, mRNA sequence
Ms.450863	BAIAP2L1 BAI1-associated protein 2-like 1
Ms.392118	KRT8C Keratin 8C
Ms.3422	FGR Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog
Ms.308050	KRT6B Keratin 6B
Ms.300279	KRT6A Keratin 6A
Ms.426704	LOC100507412 Uncharacterized LOC100507412
Ms.328989	KRT77 Keratin 77
Ms.303302	TGFB1 Transforming growth factor, beta-induced, 68kDa
Ms.316111	DCTN1 Dynactin 1
Ms.311605	ANXA2 Annexin A2
Ms.344077	GAPDH Glyceraldehyde-3-phosphate dehydrogenase
Ms.428276	CDK14 Cyclin-dependent kinase 14
Ms.353750	NPM1 Nucleophosmin (nucleolar phosphoprotein B23, numatrin)
Ms.381163	HMGB2 High mobility group nucleosomal binding domain 2
Ms.511145	FXR1 FXR1A, 1 (alpha)

MASCOT : Sequence Databases © 2007-2014 Matrix Science **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.

1. gi12787520 Mass: 14632 Score: 419 Matches: 8(9) Sequences: 7(4) mpAI: 0.62 Frame: 2 BX456308 Homo sapiens PLACENTA Homo sapiens cDNA clone C500502Y021 5-PRIME, mRNA sequence <input type="checkbox"/> Check to include this hit in error tolerant search or archive report											
Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide	
42	462.6807	923.3468	923.5116	-0.1649	0	33	35	2	U	R.FPYVALSK.T	
45	567.6567	1133.2987	1133.5499	-0.2511	0	45	4.4	1	U	R.GNEVTSVGNR.A + Oxidation (M)	(M)
81	614.2001	1226.3856	1226.6329	-0.2473	0	28	1.9e+02	2	U	K.LGPELPLASDR.F + Oxidation (M)	(M)
100	653.2101	1304.4057	1304.6837	-0.2780	0	87	0.00023	1	U	K.GNFQTIGLSAAR.F	
125	726.1806	1450.3465	1450.6477	-0.3011	0	69	0.015	1	U	R.NWYSDADVASAR.Q	
228	975.8100	1949.6055	1950.0245	-0.4190	0	86	0.00019	1	U	R.NLIFLGQGVSTVTAR.I + Oxidation (M)	
241	1001.2027	2000.3908	2000.8058	-0.4150	0	(65)	0.022	1	U	R.BGTDPPEYDDYSQGGTR.L + Oxidation (M)	
247	657.8046	2000.3919	2000.8058	-0.4139	0	73	0.004	1	U	R.NLTPDPPEYDDYSQGGTR.L + Oxidation (M)	

1. Ms_333502 Score: 702 Matches: 13(7) Sequences: 12(6) ALPPL2 Alkaline phosphatase, placental-like 2 <input type="checkbox"/> Check to include this hit in error tolerant search or archive report											
Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide	
27	462.6807	923.3468	923.5116	-0.1649	0	33	35	2	U	R.FPYVALSK.T	
41	517.1760	1032.3375	1032.5604	-0.2229	0	68	0.018	1	U	R.GSSIFGLAPGK.A	
65	567.6567	1133.2987	1133.5499	-0.2511	0	45	4.4	1	U	R.GNEVTSVGNR.A + Oxidation (M)	(M)
100	653.2101	1304.4057	1304.6837	-0.2780	0	87	0.00023	1	U	K.GNFQTIGLSAAR.F	(M)
125	726.1806	1450.3465	1450.6477	-0.3011	0	69	0.015	1	U	R.NWYSDADVASAR.Q	
170	864.2888	1726.5629	1726.9294	-0.3664	0	43	4	1	U	K.AYTULLVGAGPGVLR.D	
204	956.2437	1910.4729	1910.8601	-0.3872	0	24	2.8e+02	3	U	R.DSTLDPSLBEHTEALR.L + 2 Oxidation (M)	
208	975.8100	1949.6055	1950.0245	-0.4190	0	86	0.00019	1	U	R.NLIFLGQGVSTVTAR.I + Oxidation (M)	
202	976.2340	1950.4534	1950.8555	-0.4021	1	27	1.3e+02	1	U	K.DGARPDVTESGESGPEYR.Q	
216	1001.2027	2000.3908	2000.8058	-0.4150	0	(65)	0.022	1	U	R.BGTDPPEYDDYSQGGTR.L + Oxidation (M)	
217	667.8046	2000.3919	2000.8058	-0.4139	0	73	0.004	1	U	R.BGTDPPEYDDYSQGGTR.L + Oxidation (M)	
245	766.2128	2295.6165	2296.1084	-0.4919	0	55	0.19	1	U	R.QQSAVPLDGETHAGEDVAVFAR.G	
253	790.2187	2367.6341	2368.1295	-0.4954	0	93	3.2e-05	1	U	R.QQSAVPLDEETHAGEDVAVFAR.G	

10. gi10801208 Mass: 17215 Score: 244 Matches: 3(2) Sequences: 3(2) mpAI: 0.62 DB462453 RIKEN full-length enriched human cDNA library, testis Homo sapiens cDNA clone HD1309HD4 5', mRNA sequence <input type="checkbox"/> Check to include this hit in error tolerant search or archive report											
Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide	
65	567.6567	1133.2987	1133.5499	-0.2511	0	45	4.4	1	U	R.GNEVTSVGNR.A + Oxidation (M)	
110	820.7283	1839.6420	1839.7703	-0.1343	0	106	2.2e-06	1	U	R.ALTEITZPQKIER.A + Oxidation (M)	
253	790.2187	2367.6341	2368.1295	-0.4954	0	93	3.2e-05	1	U	R.QQSAVPLDEETHAGEDVAVFAR.G	

When we look at individual hits in the report, we see the benefits of UniGene mapping. Here we have four hits from the EST search. The entry names give no clue as to the protein function. However, when we look at the UniGene report, we find that these matches all belong to the same gene, for alkaline phosphatase.

The screenshot displays the Mascot search results interface. At the top, there's a 'Peptide Summary Report' tab and a 'RAFT - 8 (Mascot Search)' tab. The URL bar shows a file path: '54.243.190.62/mascot/cgi/master_results_2.pl?file=.%2Fdata%2F20140905%2F001270.dat_ignorescorebelow=0_minpeplen=7_prefertaxonomy'. Below the URL bar, there's a 'Threshold (0): 0' input field and a 'Cut' button.

The main results section shows a table with columns: 'Score', 'Mass', 'Matches', 'Sequences', and 'empAI F'. The first row is highlighted with a score of 229 and mass of 34632. Below this, there's a section titled '▼ 14 peptide matches (14 non-duplicate, 0 duplicate)' with an 'Auto-fit to window' option. This section contains a detailed table of peptide matches with columns: 'Query', 'Protein', 'Observed', 'Mr (exptl)', 'Mr (calc)', 'Delta M', 'Score', 'Expect', 'Rank', 'Peptide', and 'Modification'. The peptides listed include R.FFTVALSF.T, R.GSIFPULAPW.A, R.GMPTIVYDNR.A + Oxidation (M), R.LQELFLAICR.F + Oxidation (M), R.GMPTIVYDNR.F, R.MWYDADYVAAAR.Q, R.ALTETIMFIDATER.A + Oxidation (M), R.AYTVLLYDNGYVYV.L, R.KLIIFLQDGMVSTVTAAR.I + Oxidation (M), R.LQELFLAICR.F + Oxidation (M), R.MUTTFEYVYDYSQOUTR.L + Oxidation (M), R.MUTTFEYVYDYSQOUTR.L + Oxidation (M), R.QQAVFLDQETSAEDVAVFAR.G, and R.QQAVFLDQETSAEDVAVFAR.G.

At the bottom, there's a summary: '► 24 subsets and intersections (286 subset proteins in total)'.

MASCOT : Sequence Databases

© 2007-2014 Matrix Science



The protein family summary does a similar job of grouping entries together, but it can only connect overlapping entries which have at least one shared peptide match, so it will sometimes fail. The other advantage of Unigene is that it gives us the more useful descriptions.

Human Genome Statistics

- 3×10^9 bases
(EST_human is $\sim 4.4 \times 10^9$ bases)
- 6×10^9 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

© 2007-2014 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 human genome.

The human genome assembly is approximately 3 billion bases, which makes it a little smaller than EST_human.

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

In the human 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 human 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!

The screenshot shows the NCBI Map Viewer interface for the Homo sapiens (human) genome. The browser window is titled "Entrez Genome view - Windows Internet Explorer" and the URL is "http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606". The page features a search bar with "on chromosome(s)" entered, and tabs for "Published", "Nucleotide", "Protein", "Genome", "Core", "Structure", "PopSet", "Taxonomy", and "Help". The main content area displays the "Homo sapiens (human) genome view" for "Build 37.1 statistics". It includes a "BLAST search the human genome" link and a "Switch to previous build" link. The chromosome maps are shown as vertical bars representing the size of each chromosome. The lineage is listed as: Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorhina, Catarrhini, Hominoidea, Homo, Homo sapiens. A note dated August 2009 mentions the release of the updated human genome reference genome assembly (GRCv37) and the addition of 9 alternate loci to the reference assembly definition. The previous version of the reference genome assembly, NCBI Build 36.3, can still be accessed for Map Viewer display and for BLAST. The page also includes a "Release Notes" link.

MASCOT : Sequence Databases


© 2007-2014 Matrix Science

MATRIX SCIENCE

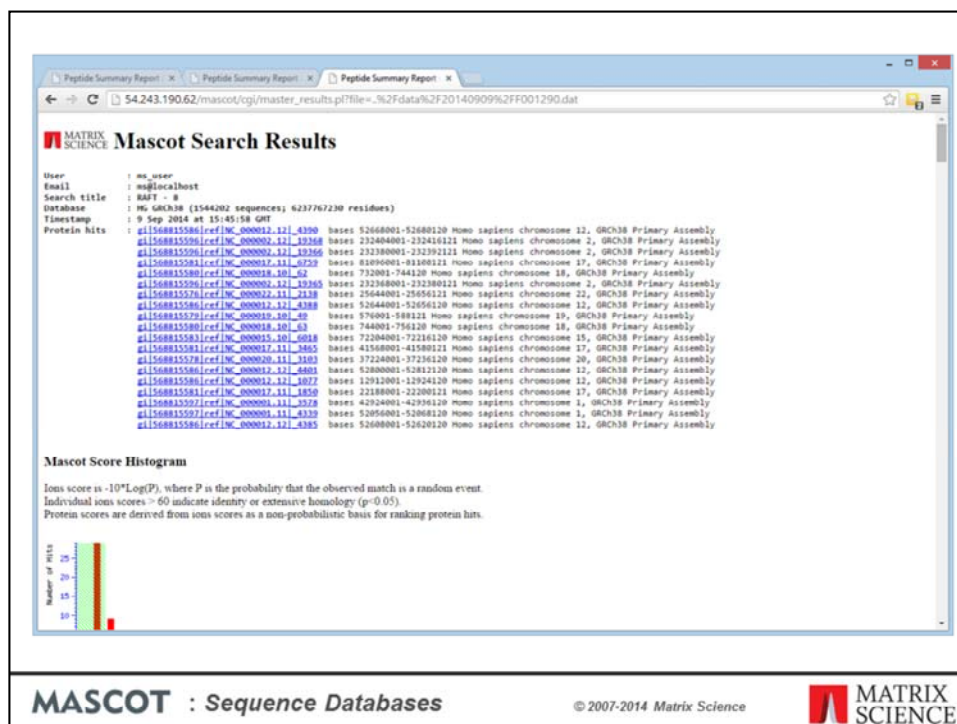
You can download the human genome sequences from NCBI.

The screenshot shows a web browser window with the address bar displaying `ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/Assembled_chromosomes/seq/`. The main content area lists 24 files, each representing a chromosome. The files are named `hs_ref_GRCh38_chr1.fa.gz` through `hs_ref_GRCh38_chr22.mfa.gz`. Each file entry includes its size in MB and a date/time stamp of `04-02-2014 00:00:00`.

File Name	Size (MB)	Date/Time
hs_ref_GRCh38_chr1.fa.gz	65.6	04-02-2014 00:00:00
hs_ref_GRCh38_chr1.mfa.gz	70.1	04-02-2014 00:00:00
hs_ref_GRCh38_chr10.fa.gz	37.9	04-02-2014 00:00:00
hs_ref_GRCh38_chr10.mfa.gz	40.5	04-02-2014 00:00:00
hs_ref_GRCh38_chr11.fa.gz	38.0	04-02-2014 00:00:00
hs_ref_GRCh38_chr11.mfa.gz	40.6	04-02-2014 00:00:00
hs_ref_GRCh38_chr12.fa.gz	37.7	04-02-2014 00:00:00
hs_ref_GRCh38_chr12.mfa.gz	40.3	04-02-2014 00:00:00
hs_ref_GRCh38_chr13.fa.gz	27.9	04-02-2014 00:00:00
hs_ref_GRCh38_chr13.mfa.gz	29.9	04-02-2014 00:00:00
hs_ref_GRCh38_chr14.fa.gz	25.7	04-02-2014 00:00:00
hs_ref_GRCh38_chr14.mfa.gz	27.5	04-02-2014 00:00:00
hs_ref_GRCh38_chr15.fa.gz	23.9	04-02-2014 00:00:00
hs_ref_GRCh38_chr15.mfa.gz	25.5	04-02-2014 00:00:00
hs_ref_GRCh38_chr16.fa.gz	22.9	04-02-2014 00:00:00
hs_ref_GRCh38_chr16.mfa.gz	24.5	04-02-2014 00:00:00
hs_ref_GRCh38_chr17.fa.gz	22.8	04-02-2014 00:00:00
hs_ref_GRCh38_chr17.mfa.gz	24.2	04-02-2014 00:00:00
hs_ref_GRCh38_chr18.fa.gz	22.0	04-02-2014 00:00:00
hs_ref_GRCh38_chr18.mfa.gz	23.5	04-02-2014 00:00:00
hs_ref_GRCh38_chr19.fa.gz	15.8	04-02-2014 00:00:00
hs_ref_GRCh38_chr19.mfa.gz	16.7	04-02-2014 00:00:00
hs_ref_GRCh38_chr2.fa.gz	68.9	04-02-2014 00:00:00
hs_ref_GRCh38_chr2.mfa.gz	73.7	04-02-2014 00:00:00
hs_ref_GRCh38_chr20.fa.gz	17.8	04-02-2014 00:00:00
hs_ref_GRCh38_chr20.mfa.gz	19.1	04-02-2014 00:00:00
hs_ref_GRCh38_chr21.fa.gz	11.1	04-02-2014 00:00:00
hs_ref_GRCh38_chr21.mfa.gz	11.9	04-02-2014 00:00:00
hs_ref_GRCh38_chr22.fa.gz	10.8	04-02-2014 00:00:00
hs_ref_GRCh38_chr22.mfa.gz	11.4	04-02-2014 00:00:00

MASCOT : Sequence Databases © 2007-2014 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



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

Peptide Summary Report x Peptide Summary Report x Peptide Summary Report x

54.243.190.62/mascot/cgi/master_results.pl?file=.%2Fdata%2F20140909%2F001290.dat

Select All Select None Search Selected Error tolerant Archive Report

1. [gi1568815586\[ref|NC_000012.11_4300\]](#) Mass: 438487 Score: 460 Matches: 9(4) Sequences: 8(4) enPAI: 0.04
bases 52668001-52680120 Homo sapiens chromosome 12, GRCh38 Primary Assembly
☐ Check to include this hit in error tolerant search or archive report


Query	Observed	Mr(expt)	Mr(calc)	Delta Miss Score	Expect	Rank	Unique	Peptide
24	487.1696	972.3246	972.5240	-0.1994	0	51	0.57	1 K.EEISELNR.V
26	590.1806	1178.3462	1178.5931	-0.2469	0	35	25	1 U --YEEIQETAGR.H
22	651.7282	1301.4419	1301.7078	-0.2659	0	63	0.035	1 U R.SLELDSDIAEVK.A
111	679.1519	1356.4092	1356.6885	-0.2792	0	79	0.002	1 U R.LNDELKQLQAG.E
112	685.1606	1392.4001	1392.7240	-0.3248	1	49	0.87	1 U R.TNAENEFVTEK.V
118	718.2301	1474.4457	1474.7416	-0.2959	0	53	0.35	1 U R.HELLQQDTSIR.T
122	738.2697	1474.5249	1474.7780	-0.2531	0	67	0.012	1 R.FLEQQMQLQTK.H
216	795.1595	2382.4566	2382.9447	-0.4880	0	80	0.0003	1 U R.GGGGGGYSGGGSSYSGGGGYSGGGGGGGGGR.G
217	1192.2367	2382.4588	2382.9447	-0.4858	0	(23)	1.5e+02	1 U R.GGGGGGYSGGGSSYSGGGGYSGGGGGGGGGR.G

2. [gi1568815596\[ref|NC_000002.11_19268\]](#) Mass: 436162 Score: 432 Matches: 7(3) Sequences: 7(3) enPAI: 0.03
bases 232404001-23246121 Homo sapiens chromosome 2, GRCh38 Primary Assembly
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta Miss Score	Expect	Rank	Unique	Peptide
42	462.6807	923.3468	923.5116	-0.1649	0	33	23	3 R.FPPVQLSE.V
65	567.6567	1133.2987	1133.5495	-0.2511	0	45	2.7	1 R.GNEVTSQVNR.A + Oxidation (N)
100	651.2101	1304.4057	1304.6837	-0.2780	0	87	0.00015	1 R.GNFTIGLSAAR.F
116	726.1806	1450.3465	1450.6477	-0.3011	0	69	0.0094	1 R.NWYSADVPASAR.Q
116	820.7283	1639.4420	1639.7763	-0.3343	0	106	1.3e-06	1 R.ALTTETISFDIAER.A + Oxidation (N)
120	864.2888	1726.5629	1726.9294	-0.3664	0	43	2.7	1 R.AYTLLYNGPGVLR.D
245	766.2128	2295.6165	2296.1084	-0.4919	0	55	0.12	1 U R.QQSAVPLDGTAGEENVFAR.G

3. [gi1568815596\[ref|NC_000002.11_19366\]](#) Mass: 432637 Score: 334 Matches: 6(4) Sequences: 5(3) enPAI: 0.03
bases 23280001-23292121 Homo sapiens chromosome 2, GRCh38 Primary Assembly
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta Miss Score	Expect	Rank	Unique	Peptide
27	469.0760	923.3468	923.5116	-0.4356	0	33	23	3 R.FPPVQLSE.V


MASCOT : Sequence Databases © 2007-2014 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

```

BLASTCDG complement(9104..9127)
/label=Q34
/colour=2
/note="Mascot match, query=34, mass=972.52, score=51, rank=1, sequence=IEISELNR"
/blastp_file="..data/20140909/P001290.dat"
/mass=972.52
/score=51
/rank=1
/translation="IEISELNR"
BLASTCDG complement(9155..9184)
/label=Q76
/colour=2
/note="Mascot match, query=76, mass=1178.59, score=35, rank=1, sequence=YEELQITAGR"
/blastp_file="..data/20140909/P001290.dat"
/mass=1178.59
/score=35
/rank=1
/translation="YEELQITAGR"
BLASTCDG complement(9379..9414)
/label=Q99
/colour=2
/note="Mascot match, query=99, mass=1301.71, score=63, rank=1, sequence=SLDLDSIIAEVK"
/blastp_file="..data/20140909/P001290.dat"
/mass=1301.71
/score=63
/rank=1
/translation="SLDLDSIIAEVK"
BLASTCDG complement(8385..8420)
/label=Q113
/colour=2
/note="Mascot match, query=113, mass=1356.69, score=75, rank=1, sequence=LNDLEDALQQAK"
/blastp_file="..data/20140909/P001290.dat"
/mass=1356.69
/score=75
/rank=1
/translation="LNDLEDALQQAK"
BLASTCDG complement(10163..10198)
/label=Q117
/colour=2
/note="Mascot match, query=117, mass=1392.72, score=49, rank=1, sequence=THAENEFTYIKK"
/blastp_file="..data/20140909/P001290.dat"
/mass=1392.72
/score=49
/rank=1
/translation="THAENEFTYIKK"
BLASTCDG complement(10679..10714)
/label=Q120

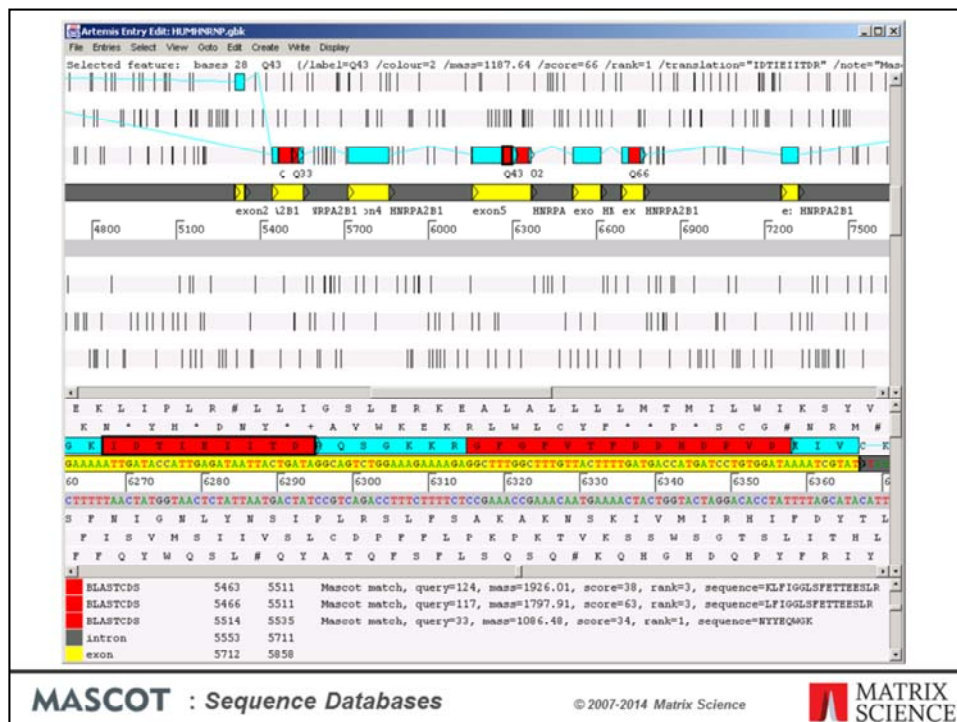
```

MASCOT : Sequence Databases © 2007-2014 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



For example, one which we find works well is Artemis, a Java based genome browser developed and distributed by the Sanger Centre.



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 stop codons. The yellow blocks are exons, while the blue blocks here are coding sequences. Individual Mascot peptide matches are shown in red. This particular gene has 8 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.

Human IPI vs. EST vs. Genome

Type of search : MS/MS Ion Search
 Enzyme : Trypsin/P
 Fixed modifications : [aCarbamidomethyl \(C\)](#)
 Variable modifications : [aAcetyl \(N-term\)](#), [aPhospho \(Y\)](#), [aPhospho \(ST\)](#), [aOxidation \(M\)](#), [aGln->pyro-Glu \(N-term Q\)](#)
 Mass values : Monoisotopic
 Protein mass : Unrestricted
 Peptide mass tolerance : ± 10 ppm ($\#^{13}\text{C} = 1$)
 Fragment mass tolerance : ± 0.6 Da
 Max missed cleavages : 1
 Instrument type : ESI-TRAP
 Number of queries : 8,797

Database	Size in residues	Average score threshold	Number of PSMs (1% FDR)
Uniprot human	3.5×10^7	34	3234
EST human	8.8×10^9	60	2017
Human genome	6.2×10^9	56	1365

MASCOT : Sequence Databases

© 2007-2014 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 iPRG2010 dataset distributed by ABRF.

There is a big drop in the number of matches between Uniprot human and EST human. The reason is mainly that EST human is a much bigger database, by more than a factor of 100. This means that the score thresholds are approx 24 higher, and we lose all the weaker matches, that had scores between 34 and 60. 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 human genome is even worse. This is not because of a still higher threshold; the database is actually smaller than EST_human. 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 much larger than 20%. The other factor is that the human 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.

Human IPI vs. EST vs. Genome

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