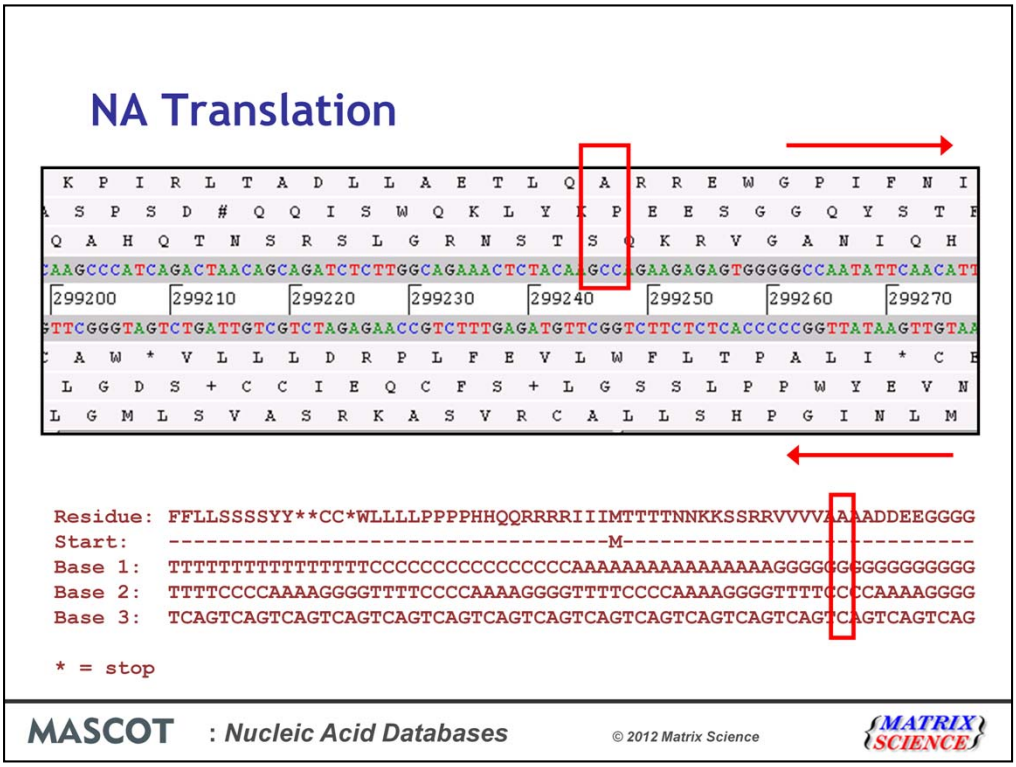
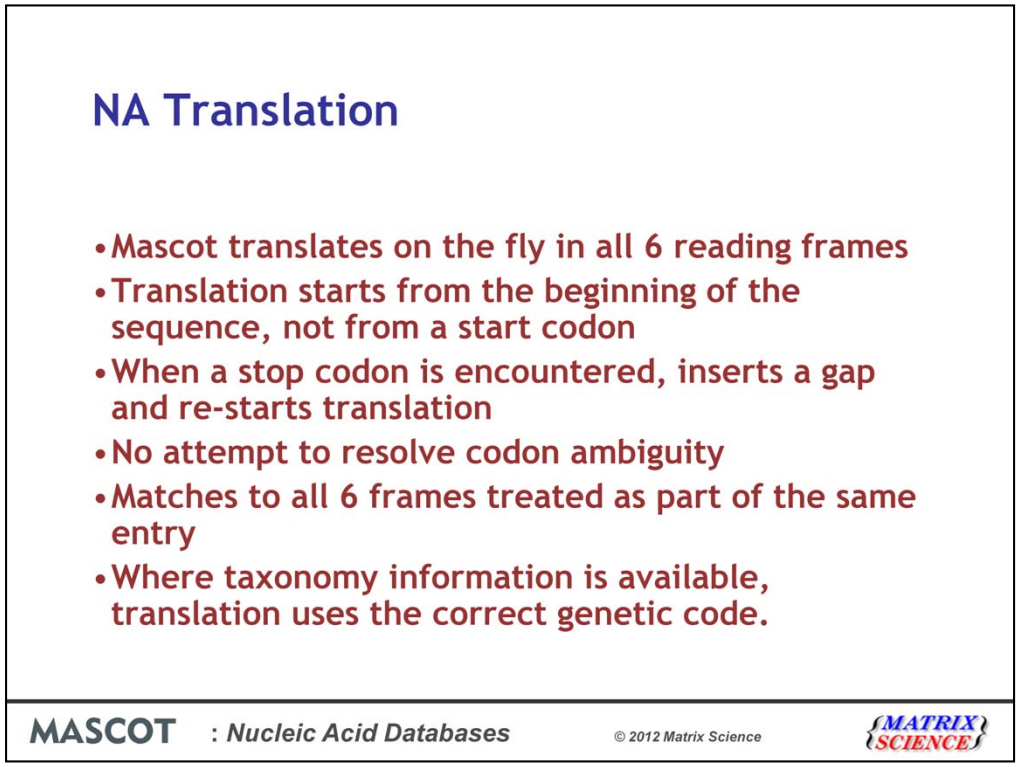
Searching Nucleic Acid Databases

MASCOT

MATRIX SCIENCE



When we search a nucleic acid databases, there is no need for you to carry out your own six frame translation. 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.



The rules for NA translation in Mascot are

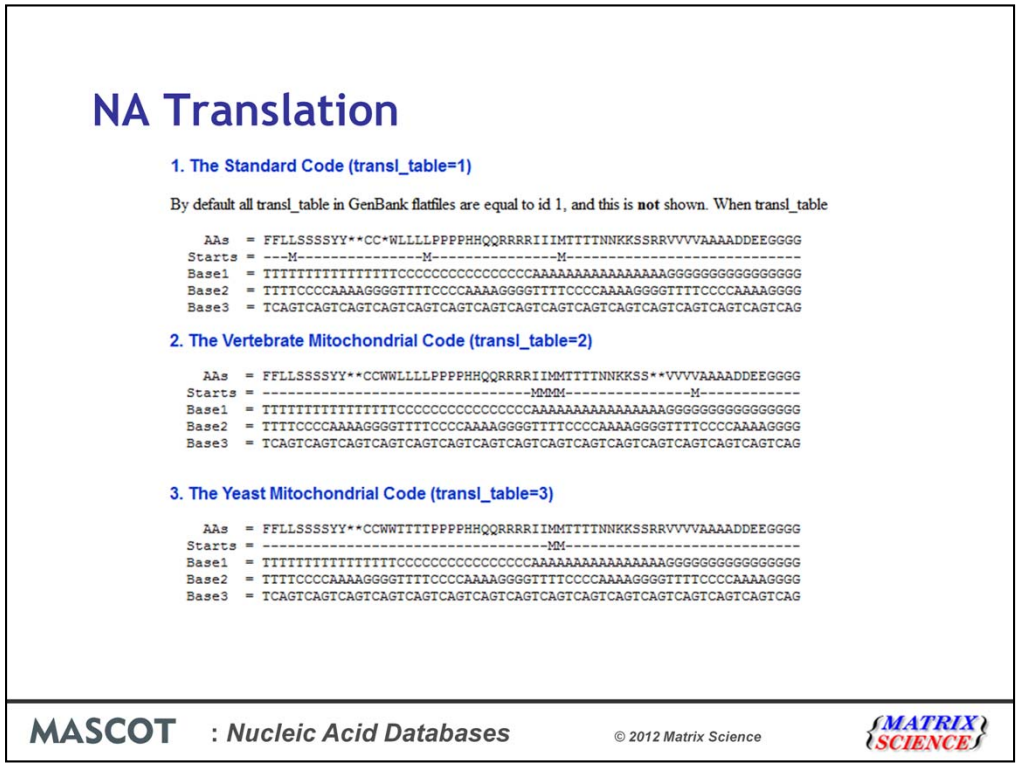
Translate the entire sequence, don't look for a start codon to begin

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.

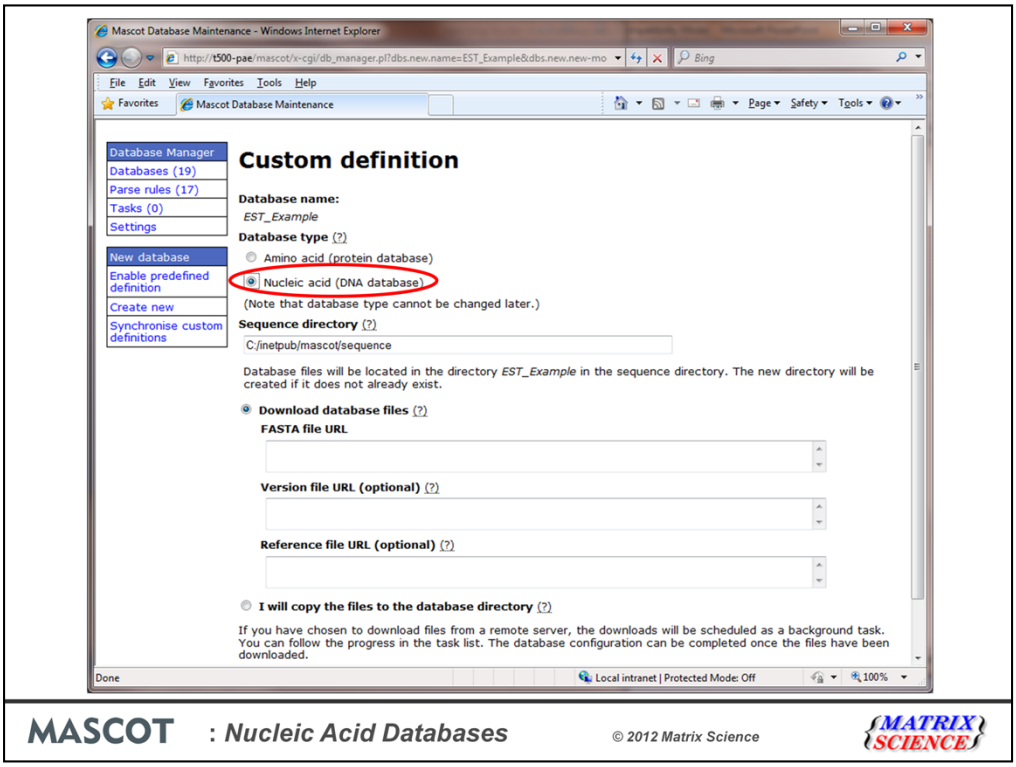
It is better to allow Mascot to search the nucleic acid database directly rather than searching a pre-translated database because Mascot will treat matches from all 6 frames as part of the same entry, so matches will remain grouped together even if there is a frame shift.

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

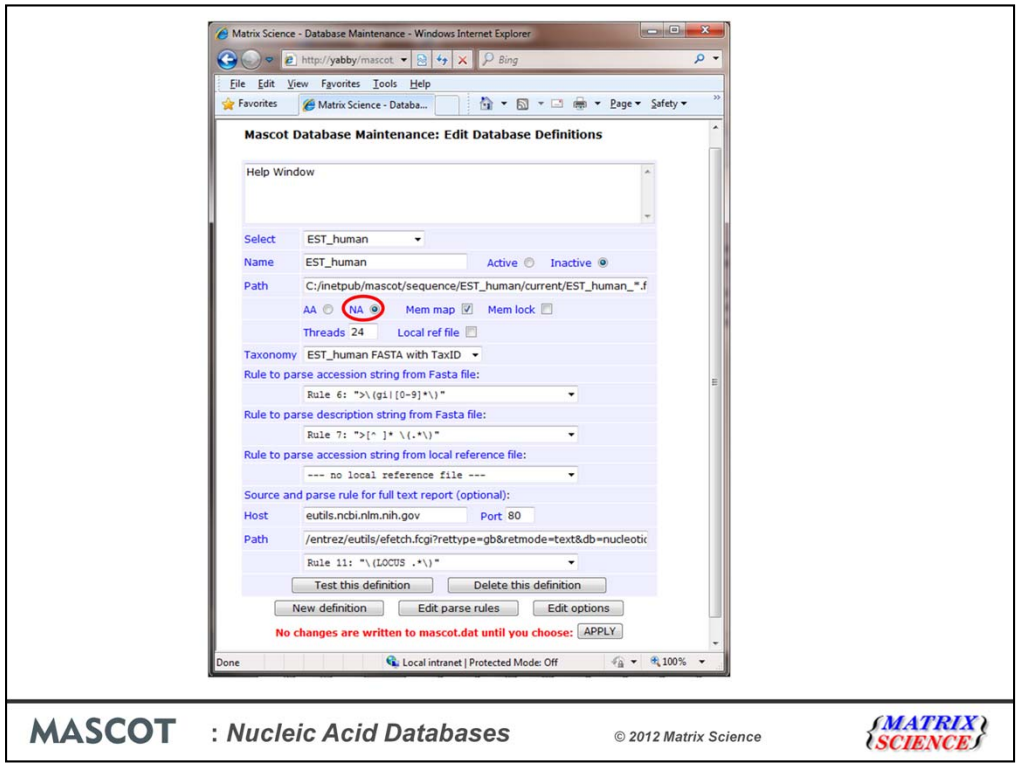


For many species, the genetic code is different for the mitochondrial and nuclear genomes. Whether to use the genomic or mitochondrial translation table for a database is set by a flag in the Mascot.dat configuration file, and applies to all the entries in that sequence database. For this reason, if you need to search both genomic and mitochondrial DNA sequences, you should not include both genomic and mitochondrial sequences in the same database, but set up a separate database for each. In Mascot 2.3 we introduced the ability to search more than one sequence database at a time, so you can still easily search both sets of sequences at the same time. Since the mitochondrial genomes are all relatively small, this will also tighten up the statistics of your search.

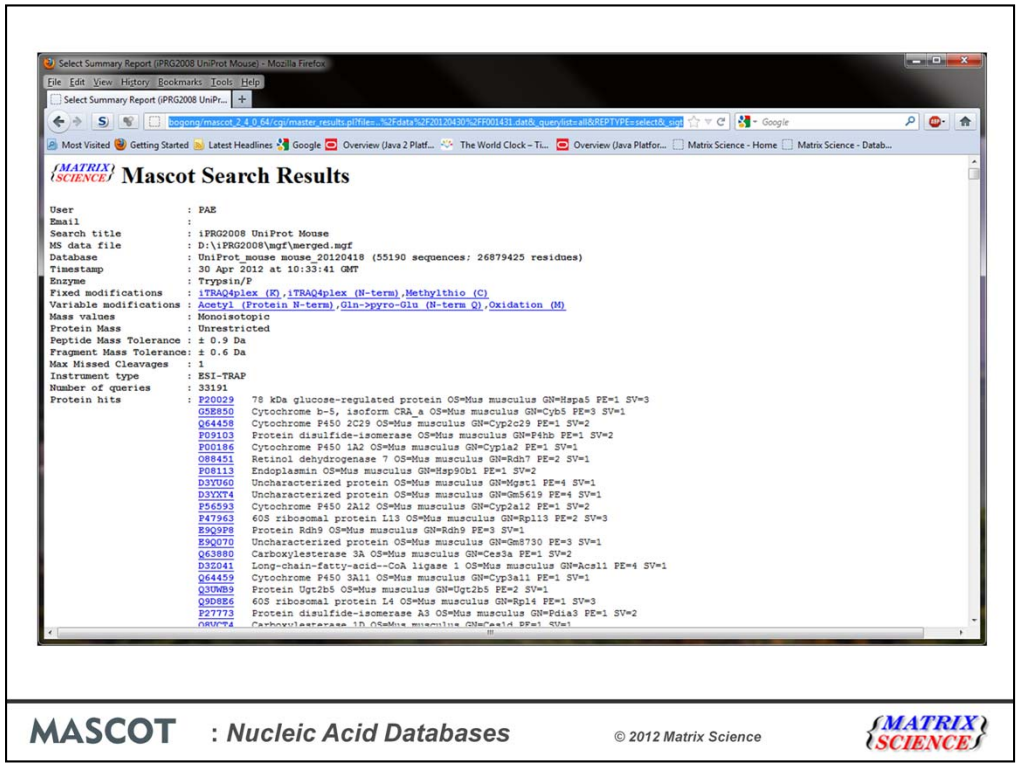
Bacterial, Archaeal and Plant plastids also use a different genetic code. Sequences from these are given a separate taxonomy id by the NCBI, so you do not need to use a separate database for these.



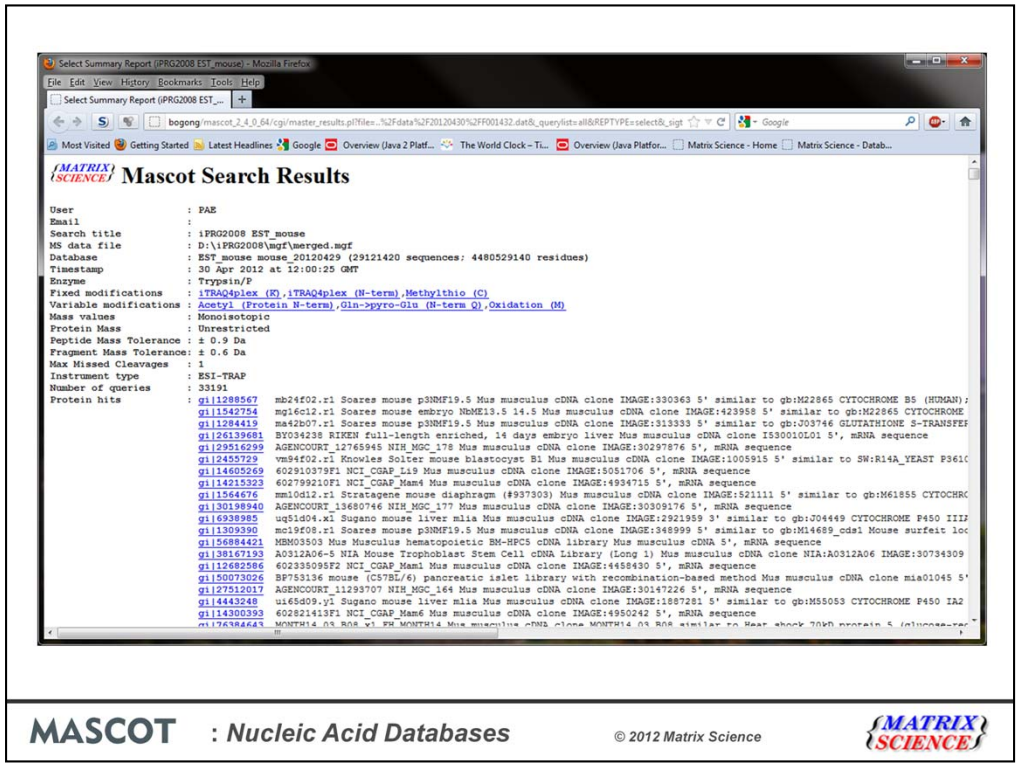
Setting up a nucleic acid database in Mascot is no different from setting up a protein database. The only thing to watch is that the database type is specified as NA. This is where you do it in Mascot 2.4 in the new Database Manager when you are setting up a custom definition.



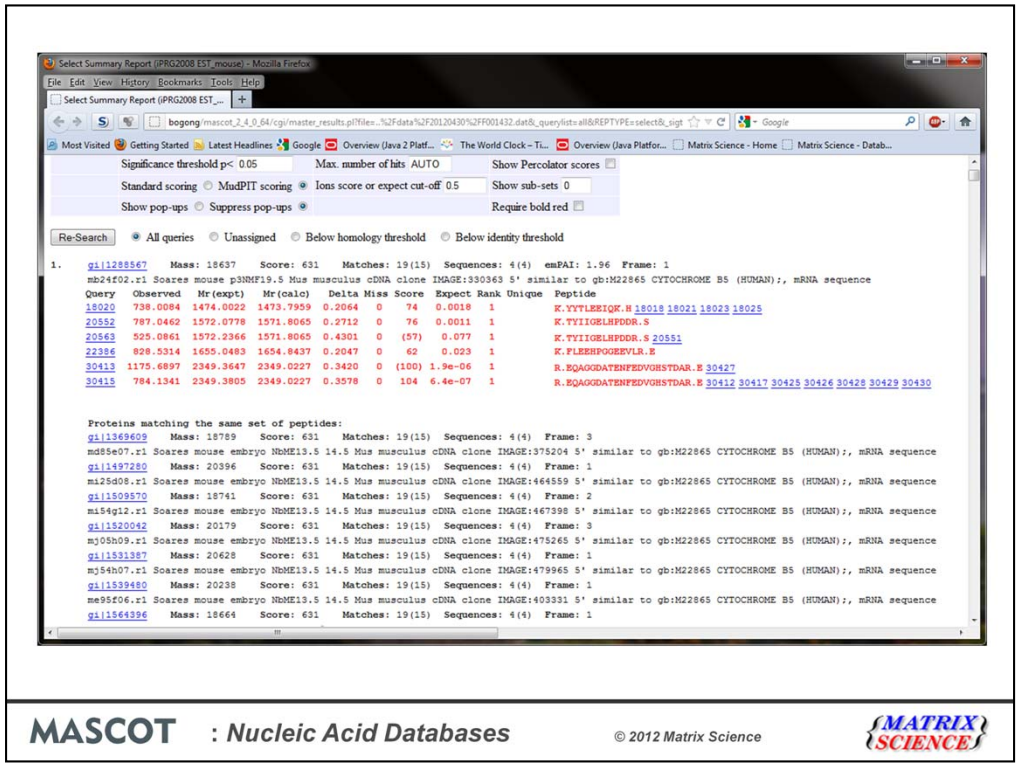
And this is where you do this in Mascot 2.3 and earlier on the database maintenance page.



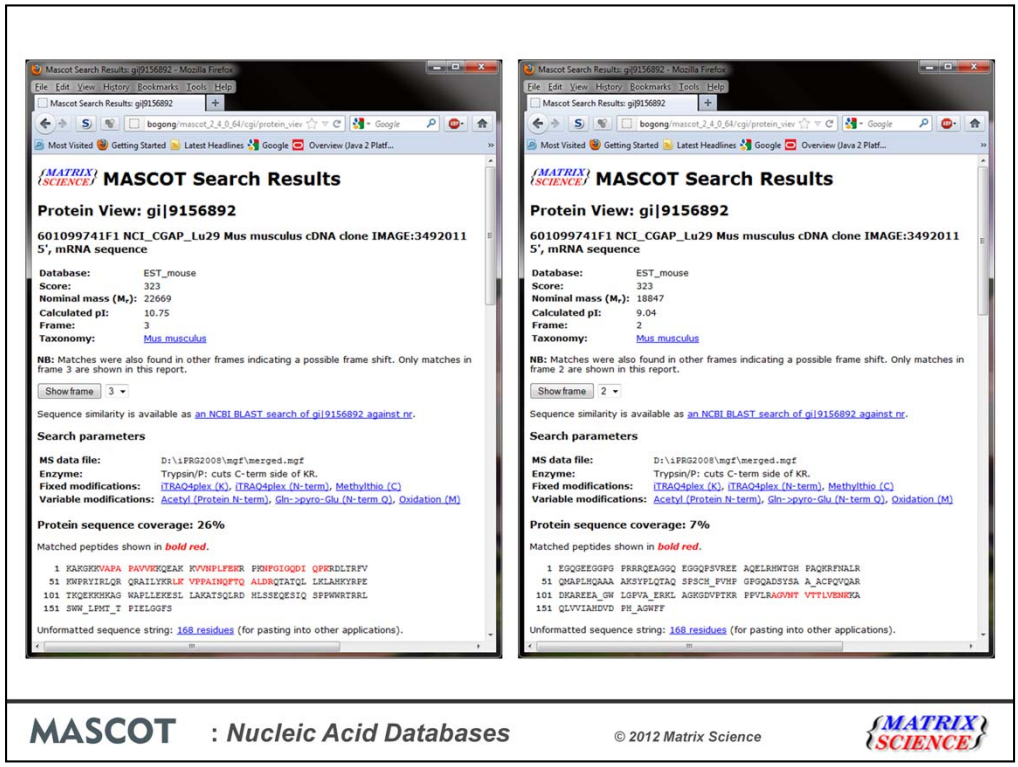
To illustrate the features of the different types of database, we searched the public iPRG2008 dataset distributed by ABRF against a protein database, UniProt mouse. We found significant matches to 438 mouse proteins.



With EST_mouse, we obtained almost the same results, just a couple of additional peptide matches. However, look at the hit-list on this report … we have a very long match list because there is a large amount of redundancy in the EST databases and unlike the protein database search, most of the matched entries don't immediately communicate which proteins have been found. I'll return to these issues later.

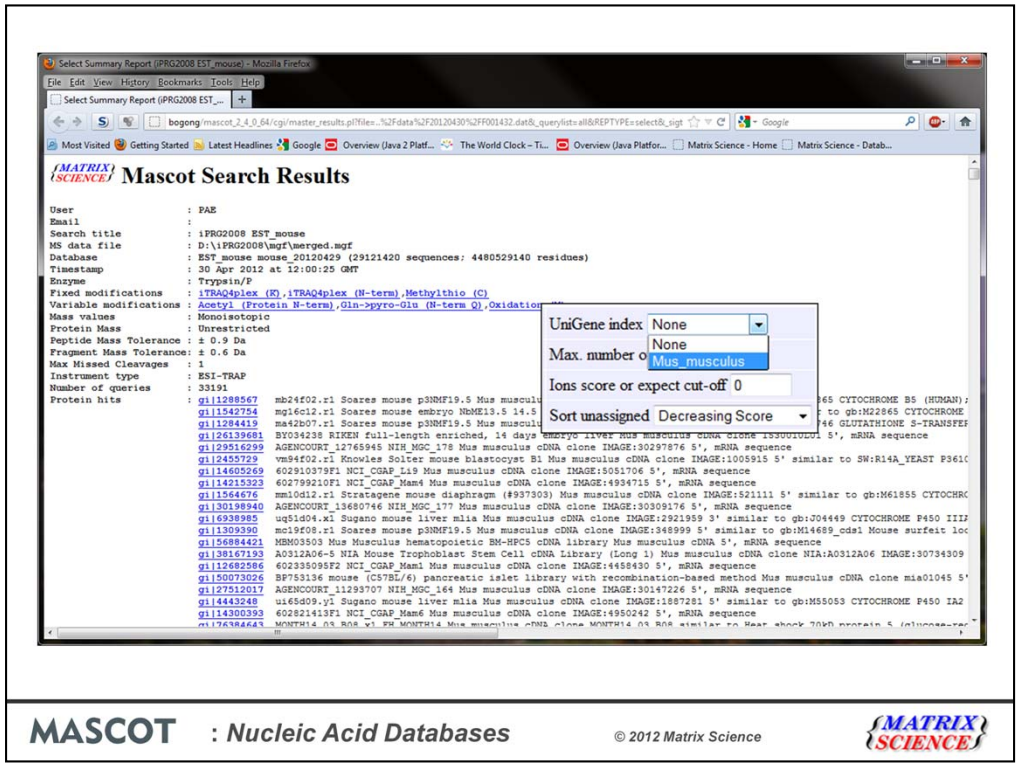


The master results report from the EST search looks pretty similar to the Uniprot Mouse search, except that the EST sequences are mostly shorter than full length proteins, so the peptide matches are more scattered. If we click on the protein accession number link



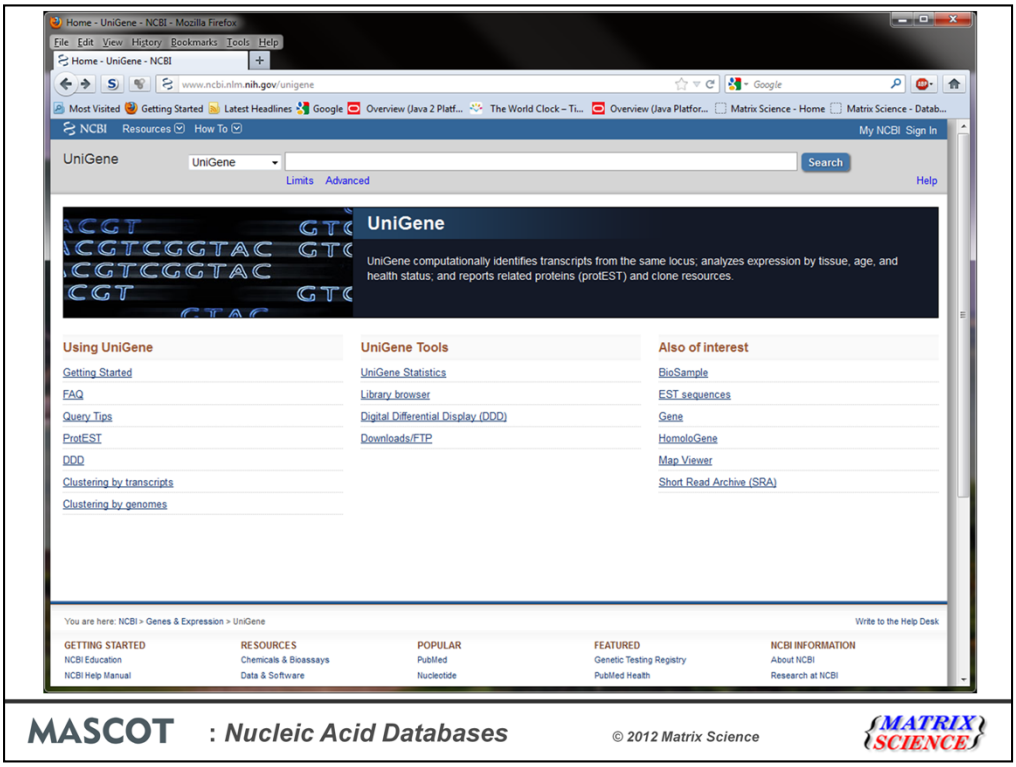
We get a protein view. This is similar to the protein view for a protein database entry, except we have drop down list for the different translation frames. For this particular entry, most of the matches have been found in reading frame 3.

But, as so often happens, there is a frame shift in this entry, and there is an additional match in frame 2.



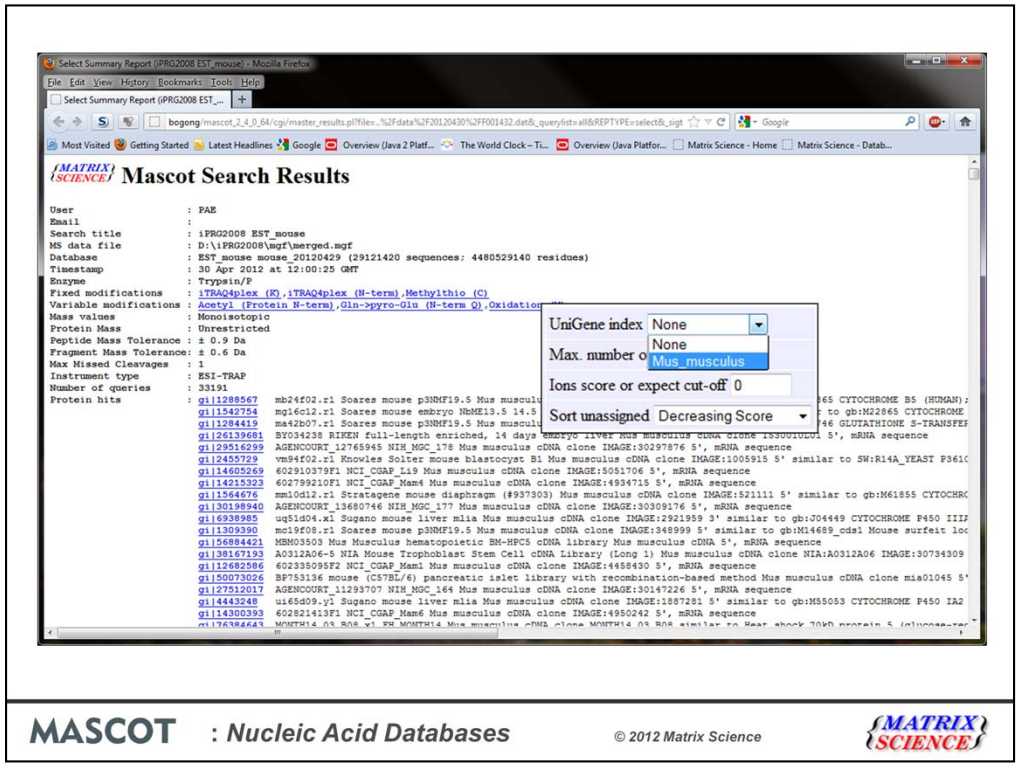
Going back to the issue of the hit list and the descriptions not saying very much. There are several problems here. One is that EST databases usually have a huge amount of redundancy, which can make for very long reports. Another problem is that the sequences tend to be short, so we don't get much grouping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

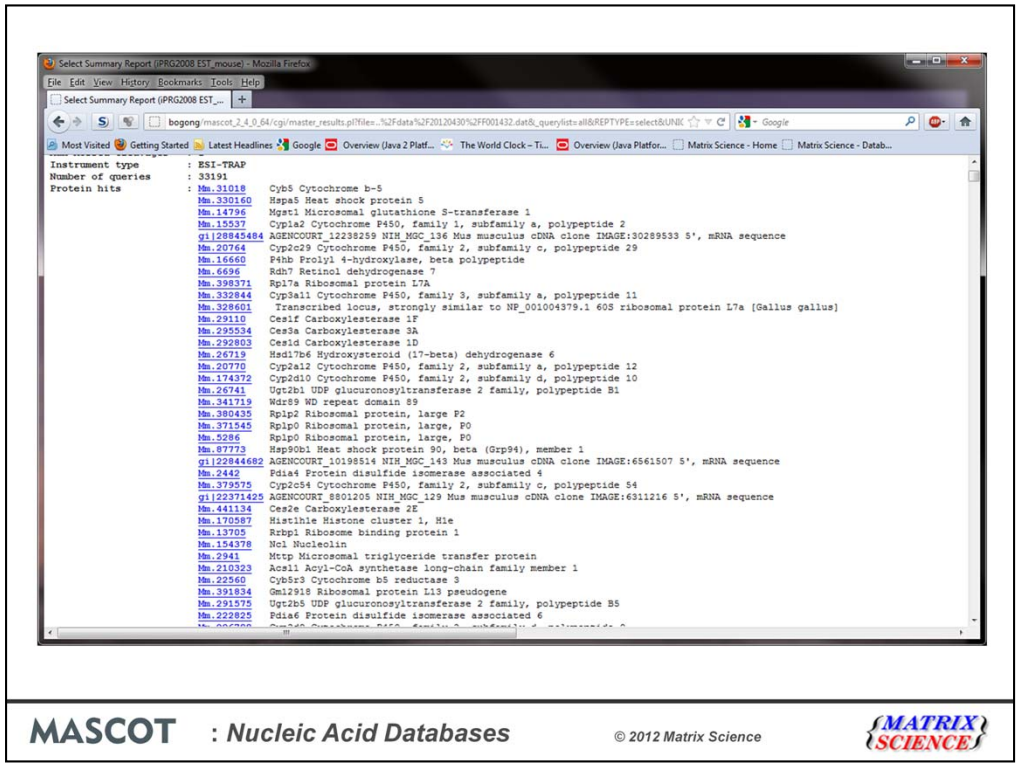


UniGene is not a sequence database, it is an index which is created by BLASTing GenBank sequences against themselves to cluster them into gene families. UniGene can be downloaded from the NCBI FTP site. In Mascot 2.4, the database manager automatically downloads and configures unigene indexes for pre-defined databases such as EST Human.

For Mascot 2.3 and earlier, downloading can be automated using the db_update.pl script. Some manual configuration of Mascot.dat is then required in order to enable the indexes.

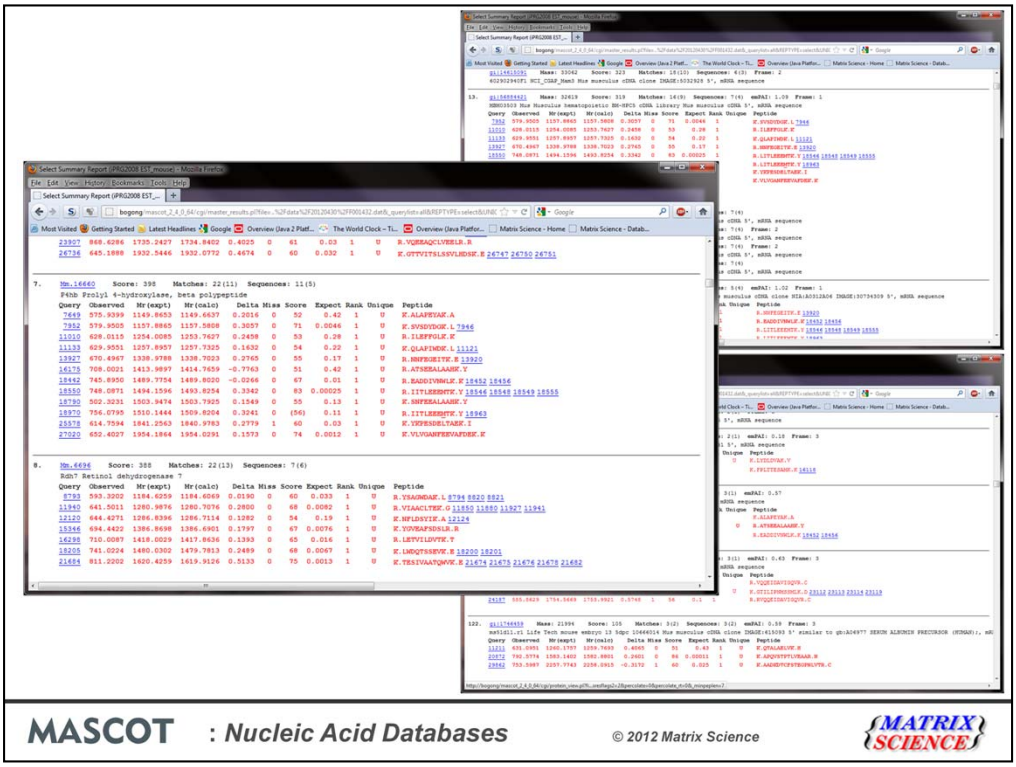


So, if Unigene is configured, we can select mouse from the drop-down list in the format controls

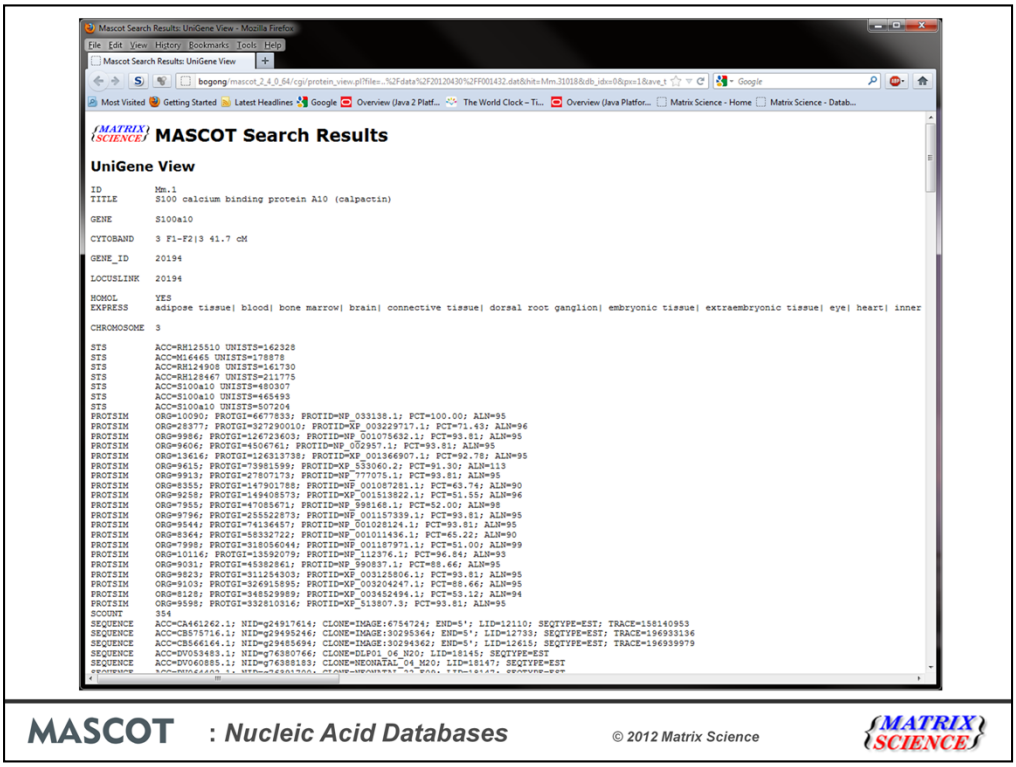


Now, using the UniGene index as a lookup table, we can transform the results of an EST search.

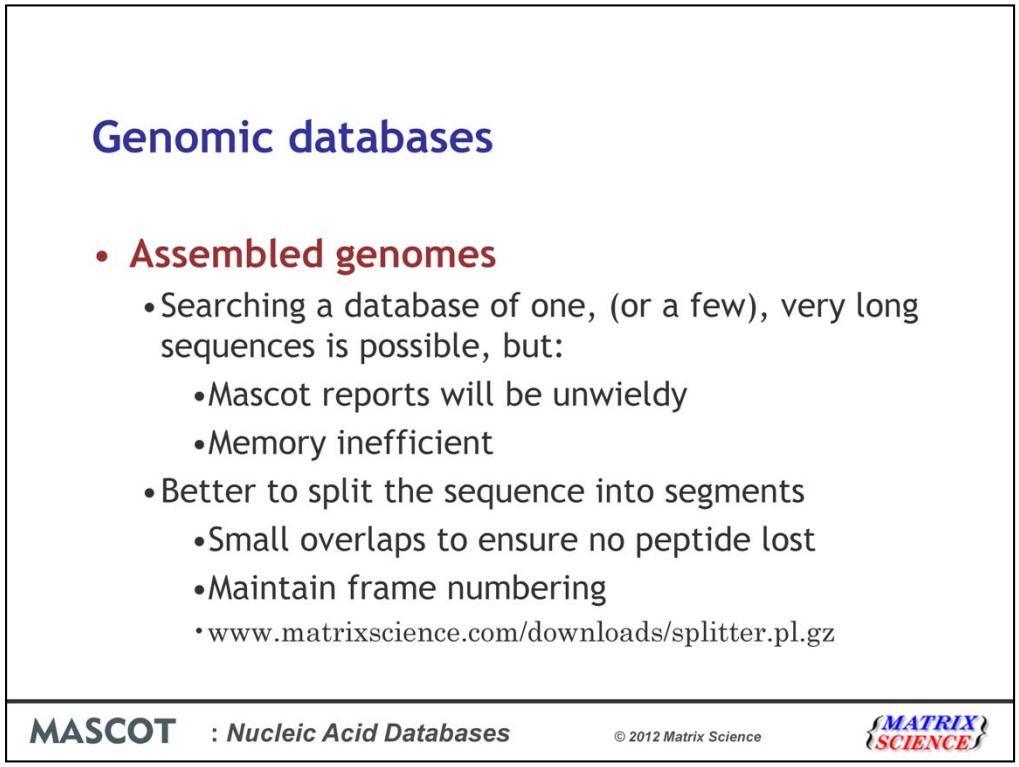
This is now a much clearer picture, very similar to the protein database result. Please remember that we are not clustering the database sequences into consensus sequences prior to searching. This could lead to matches being missed. UniGene is being used after the search, to map one set of accessions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



When we look at individual hits in the report, we see the benefits of UniGene mapping. Here we have two hits from the EST search – hits 13 and 120. They don't have any shared peptide matches, and the entry names also 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 Prolyl 4-hydroxylase.



When you click on the accession number link of a unigene filtered report, you get full details for that particular gene family.



We can also perform MS/MS searches on the raw genomic sequence data.

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

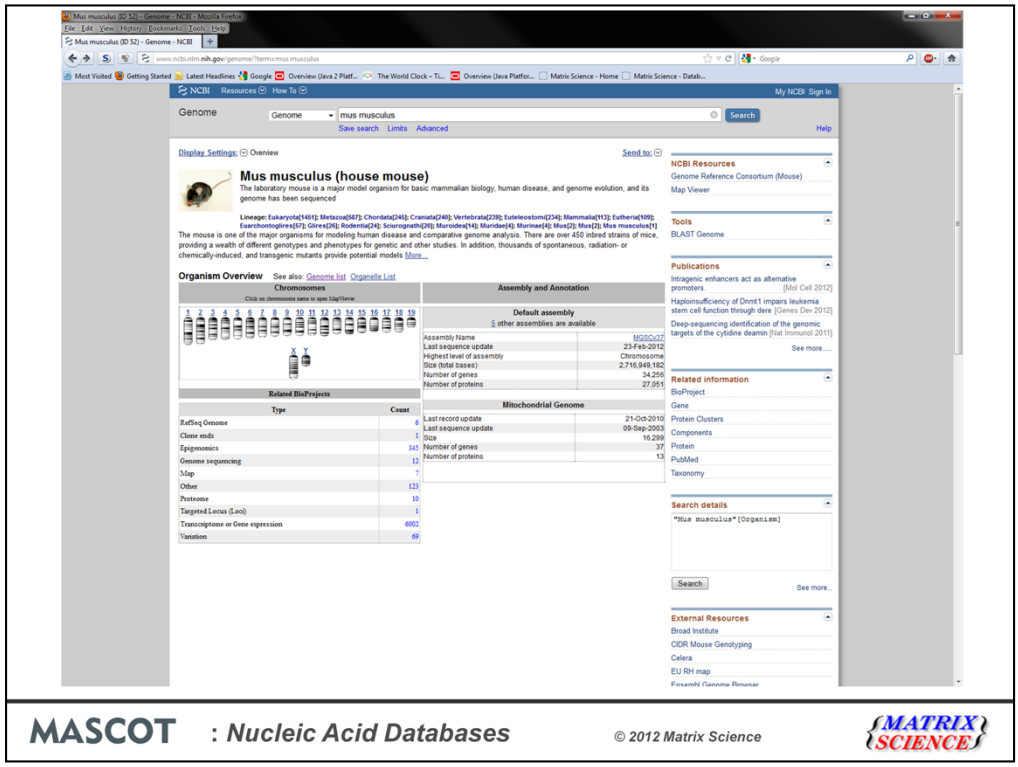
For example, the longest human chromosome is chromosome 1 with 285 million base pairs

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

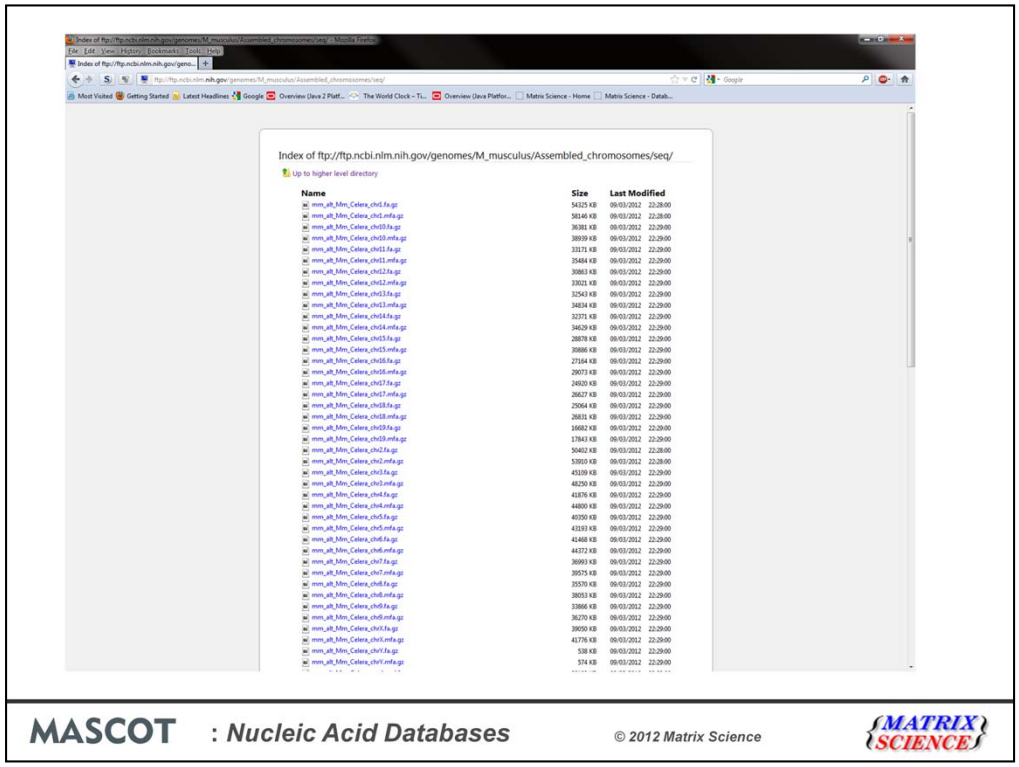
Mascot requires a significant memory overhead to manipulate such long sequences, which means that unless you have a very large amount of RAM, the search is going to be using virtual memory ... i.e. swapping out to disk ... and run relatively slowly.

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

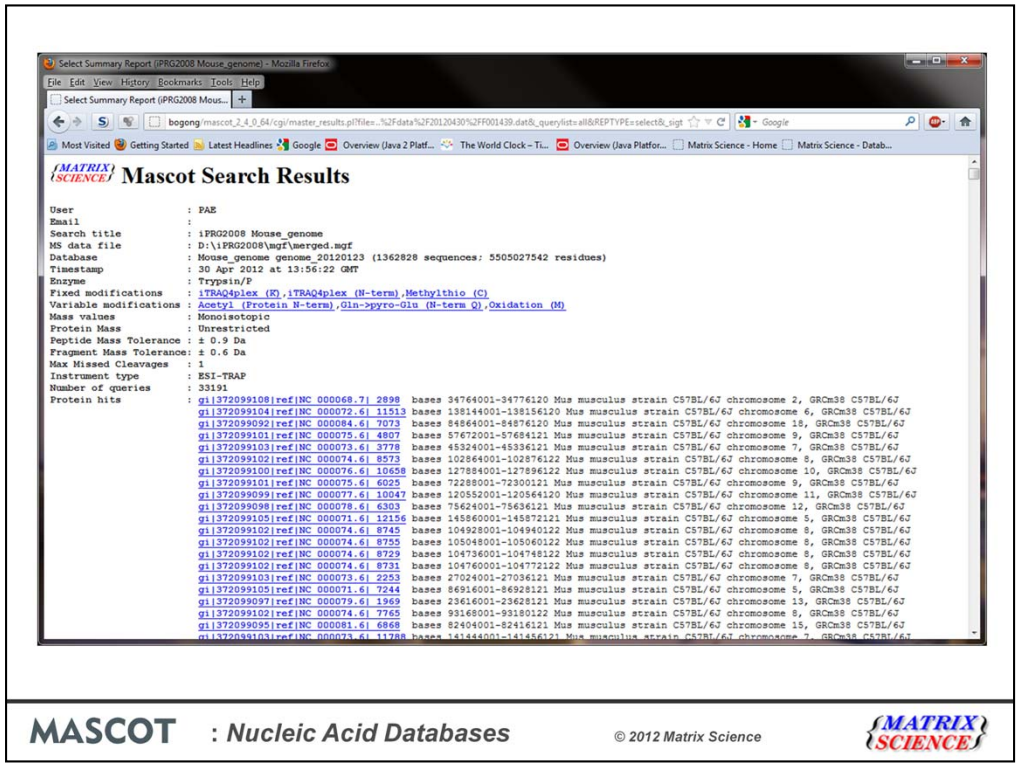
For efficient searching and reporting, the genomic DNA needs to be chopped into shorter sequences, with small overlaps to ensure no peptides are lost because they span a boundary. This is not a completely trivial task if you want to maintain the original forward and reverse frame numbering from chunk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



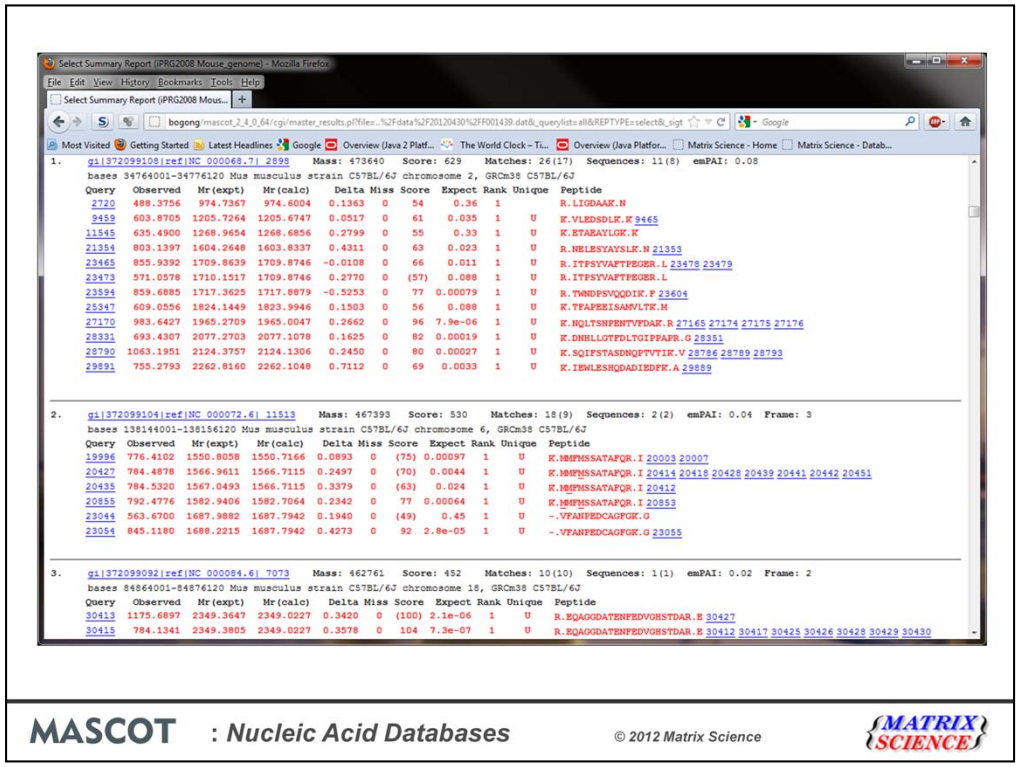
You can download the mouse genome sequences from the NCBI.



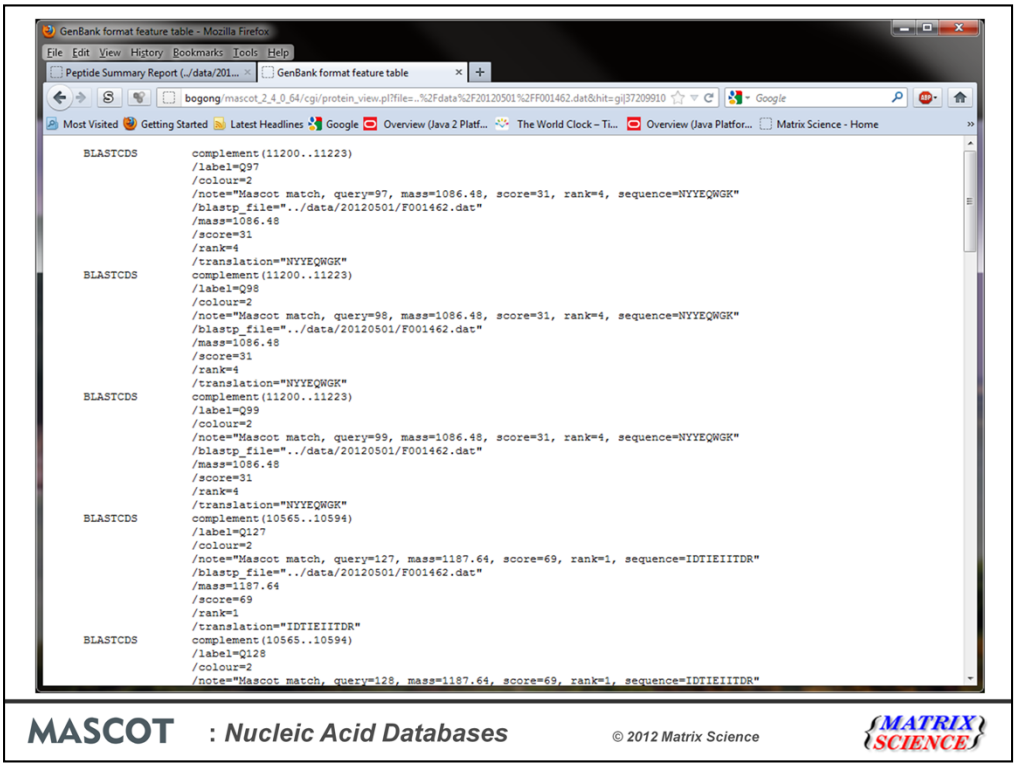
So for the mouse genome, we chose the assembled chromosomes - 21 files (using the GRCm38 reference assembly). Although you could search this as a 21 entry database, as I just mentioned this is not memory efficient, so we used the splitter script to split the chromosome sequences into overlapping segments of 12 kb



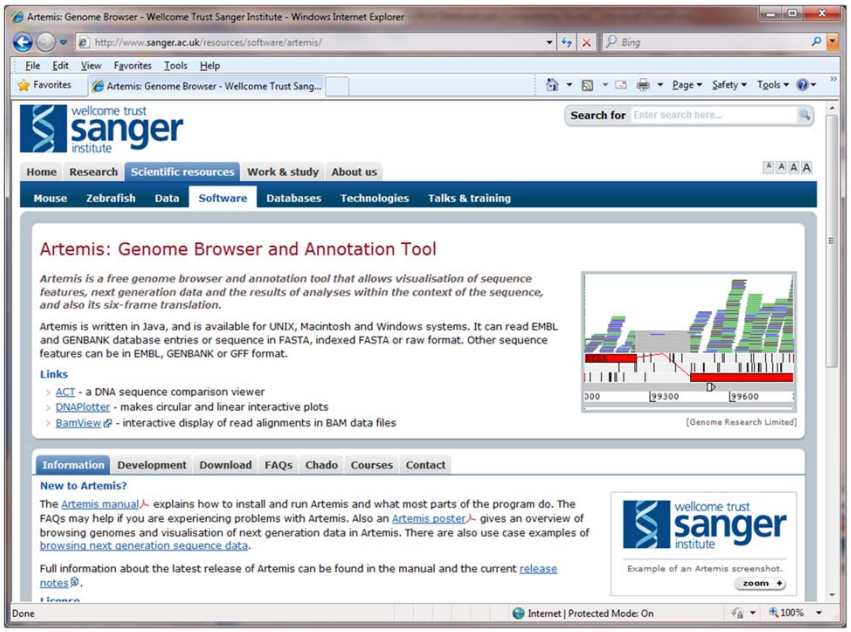
This is the result of searching our data against the mouse genome assembly. If you thought that most of the EST_mouse entry titles were uninformative, how much worse is this?



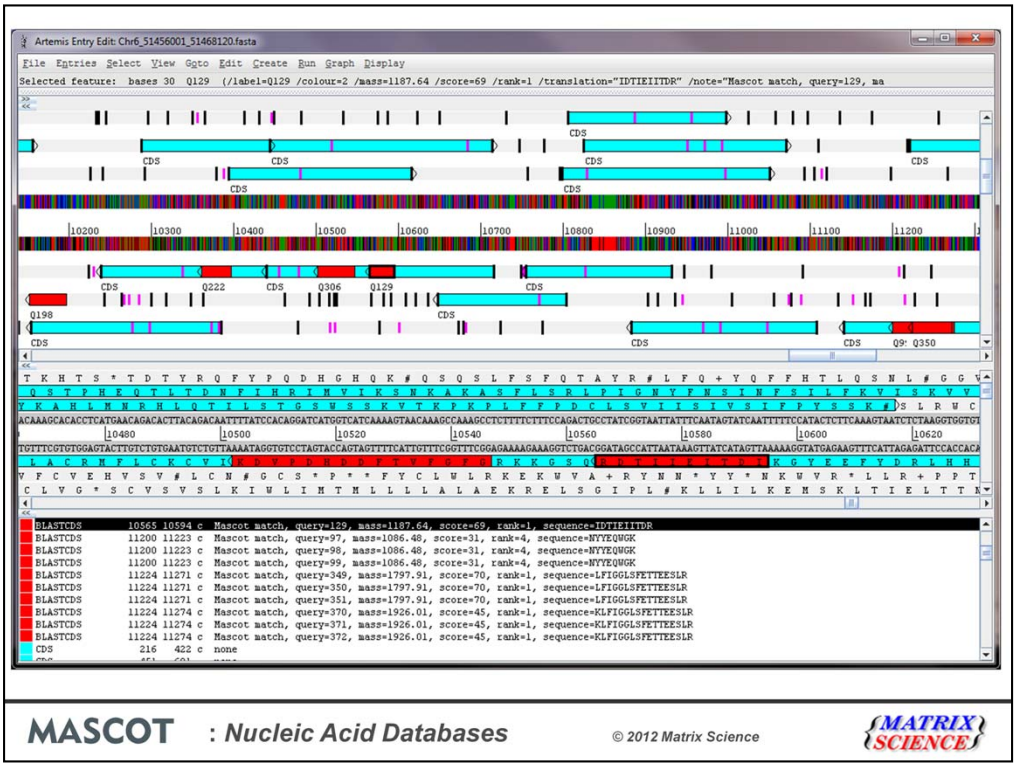
If you click on an accession number link, for a protein view report, you can get either the standard protein view report or an alternative



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.

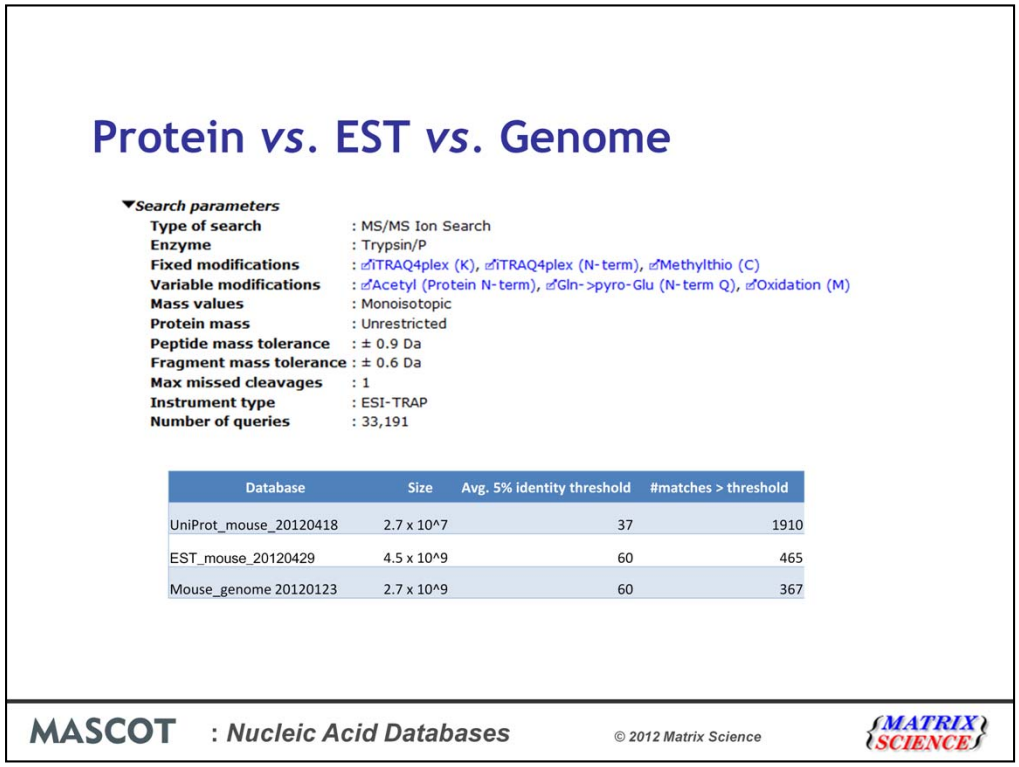


Here's the result of reading the feature table containing the Mascot peptide matches into Artemis. In the upper third, we have a low resolution view. This can be zoomed out to show an entire sequence as a single strip. We have the forward and complementary DNA strands, and the 6 frame translation. The black vertical bars are stop codons, the purple bars are start codons. The the blue blocks are open reading frames. Individual Mascot peptide matches are shown in red. This particular gene has 5 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.

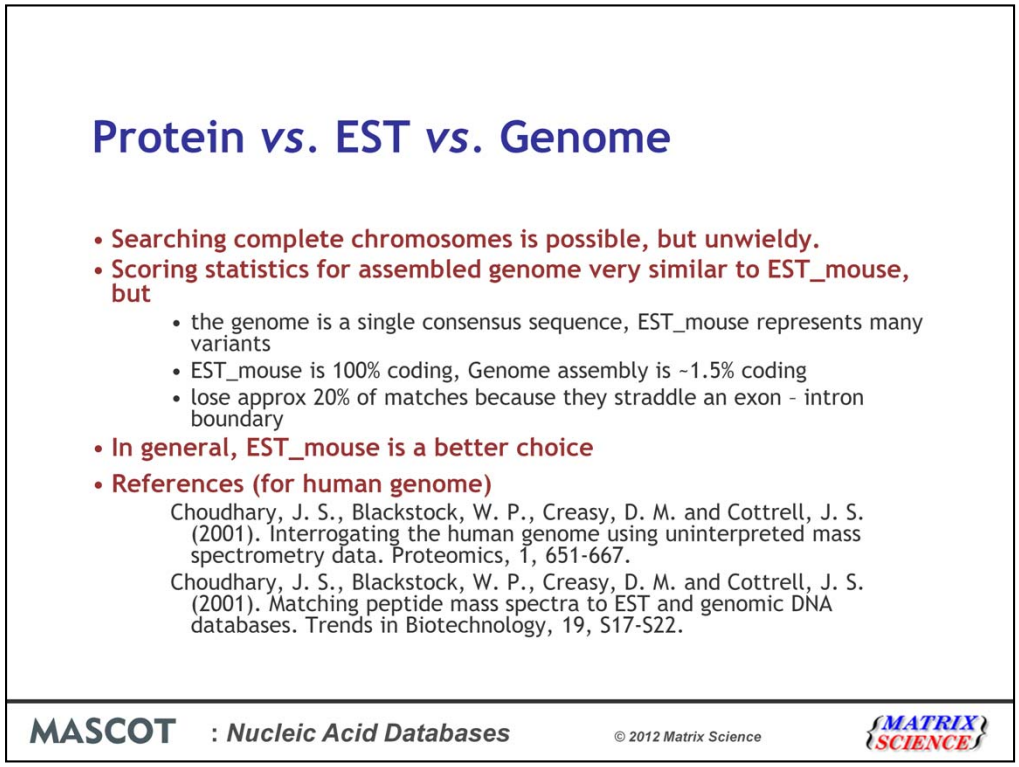
Not only does this allow us to zoom and pan around these extremely long sequences, it also allows us to view the peptide matches found by Mascot in the context of all the existing annotations. This gives us a powerful way to present the results of MS based searching complete genomes.



All well and good, but which database gives the most matches?

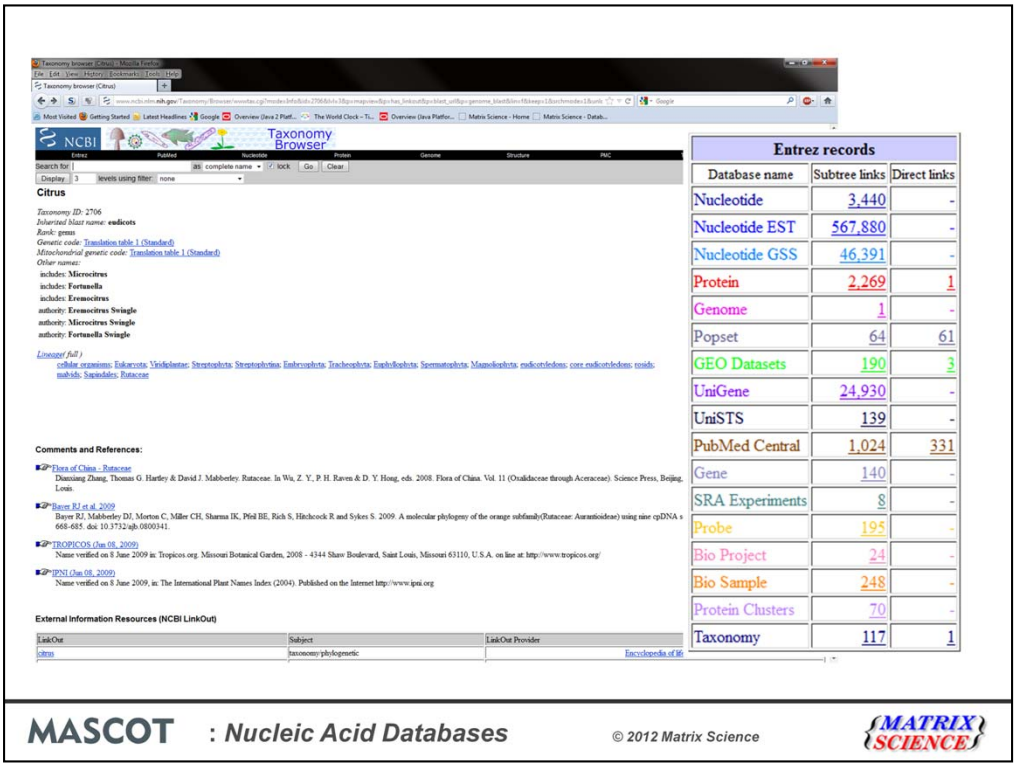
As you can see in the table, there is a big drop in the number of matches between Uniprot_mouse and EST_mouse. The reason is mainly that EST_mouse is a much bigger database, by more than a factor of 100. This means that the score thresholds are approx 23 higher, and we lose all the weaker matches, that had scores between 37 and 60. Yes, there may be additional matches in EST, not found in Uniprot mouse, 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; the database is similar in size to EST_mouse. 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 approximately 21%. The other factor than can affect this 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.

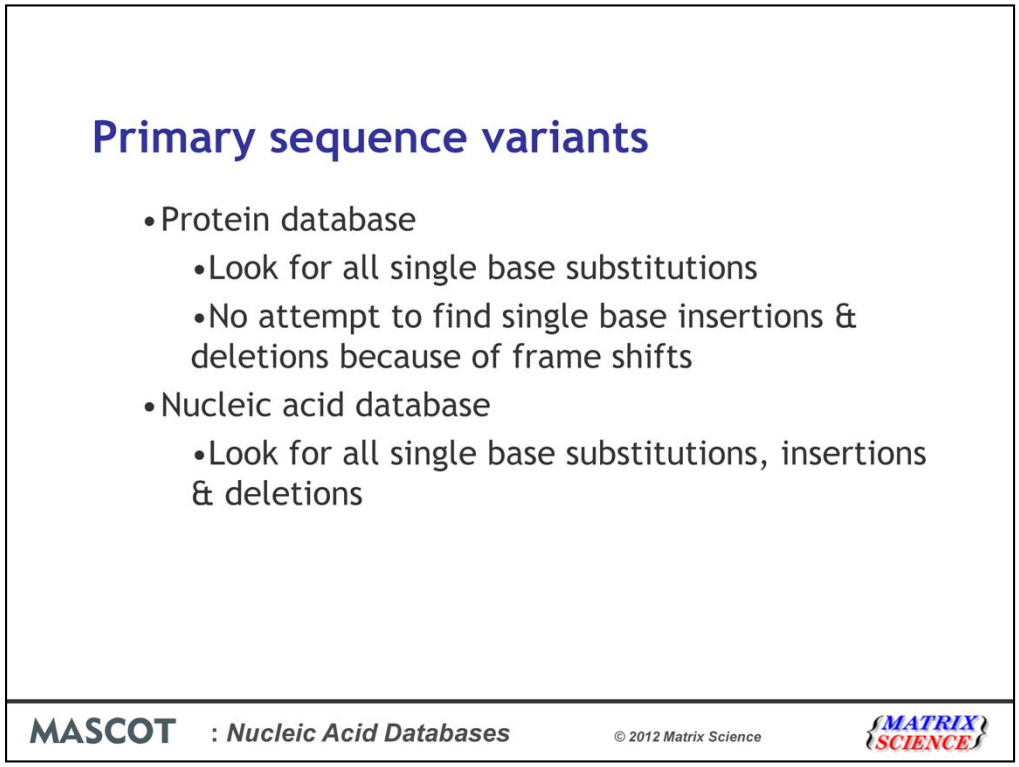


So, these are our conclusions for the mouse genome, and the same considerations probably hold for other large mammalian genomes (and certainly for the human genome).

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

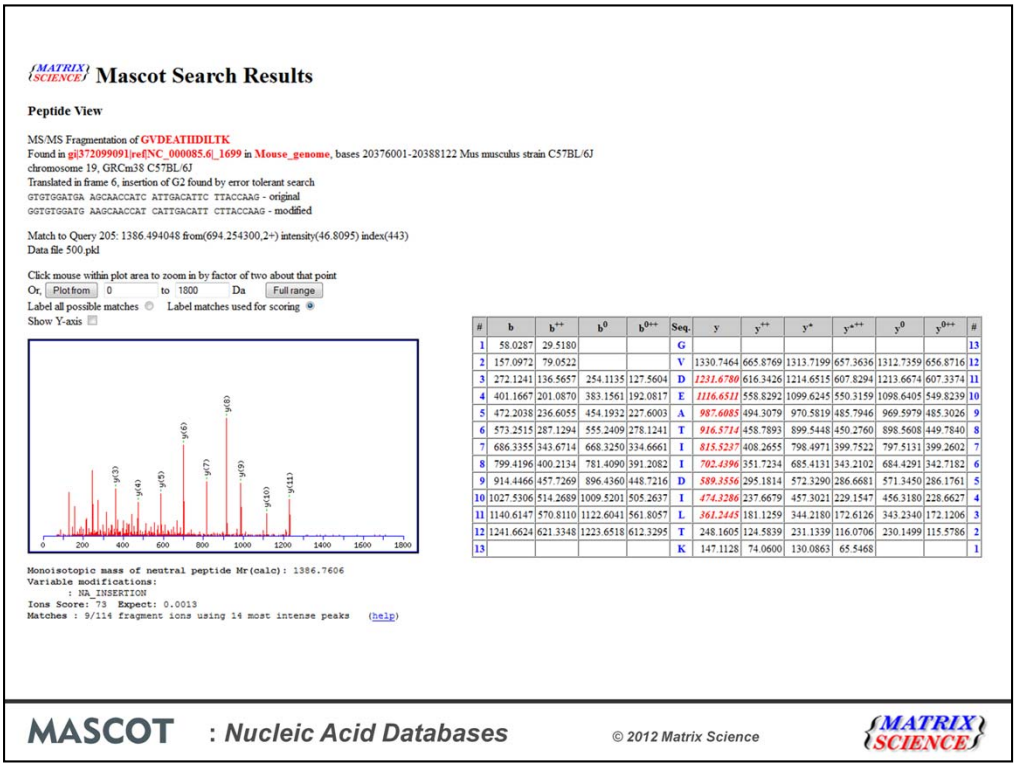


Here we're looking at the overall numbers for all the species of the genus Citrus in Genbank – and as you can see we have very few protein entries, but some genome data from a few species and reasonably large number of ESTs. So if you were working on a Citrus species, you would probably need to be searching a nucleic acid database.



An important reason for why you may wish to search against a nucleic acid database is to look for sequence variants, such as single nucleotide polymorphisms (SNPs) or sequencing errors using the error tolerant search in Mascot.

For a protein database, we can't look for the consequences of inserted or deleted bases, because these give rise to frame shifts, and the entire sequence changes from that point on, but if we're searching a nucleic acid database, we can look for all single base substitutions, insertions and deletions.



Mascot Search Results

Peptide View

MS/MS Fragmentation of **GVDEATIIDILTK**
Found in **gi|372099091|ref|NC_000085.6|_1699** in **Mouse_genome**, bases 20376001-20388122 Mus musculus strain C57BL/6J chromosome 19, GRCm38 C57BL/6J
Translated in frame 6, insertion of G2 found by error tolerant search
GTGTGGATGA AGCAACCATC ATTGACATTC TTACCAAG - original
GGTGTGGATG AAGCAACCAT CATTGACATT CTTACCAAG - modified

Match to Query 205: 1386.494048 from(694.254300,2+) intensity(46.8095) index(443)
Data file 500.pkl

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 0 to 1800 Da Full range
Label all possible matches Label matches used for scoring
Show Y-axis

#	b	b^{++}	b^0	b^{0++}	Seq.	y	y^{++}	y^*	y^{*++}	y^0	y^{0++}	#
1	58.0287	29.5180			G							13
2	157.0972	79.0522			V	1330.7464	665.8769	1313.7199	657.3636	1312.7359	656.8716	12
3	272.1241	136.5657	254.1135	127.5604	D	*1231.6780*	616.3426	1214.6515	607.8294	1213.6674	607.3374	11
4	401.1667	201.0870	383.1561	192.0817	E	*1116.6511*	558.8292	1099.6245	550.3159	1098.6405	549.8239	10
5	472.2038	236.6055	454.1932	227.6003	A	*987.6085*	494.3079	970.5819	485.7946	969.5979	485.3026	9
6	573.2515	287.1294	555.2409	278.1241	T	*916.5714*	458.7893	899.5448	450.2760	898.5608	449.7840	8
7	686.3355	343.6714	668.3250	334.6661	I	*815.5237*	408.2655	798.4971	399.7522	797.5131	399.2602	7
8	799.4196	400.2134	781.4090	391.2082	I	*702.4396*	351.7234	685.4131	343.2102	684.4291	342.7182	6
9	914.4466	457.7269	896.4360	448.7216	D	*589.3556*	295.1814	572.3290	286.6681	571.3450	286.1761	5
10	1027.5306	514.2689	1009.5201	505.2637	I	*474.3286*	237.6679	457.3021	229.1547	456.3180	228.6627	4
11	1140.6147	570.8110	1122.6041	561.8057	L	*361.2445*	181.1259	344.2180	172.6126	343.2340	172.1206	3
12	1241.6624	621.3348	1223.6518	612.3295	T	248.1605	124.5839	231.1339	116.0706	230.1499	115.5786	2
13					K	147.1128	74.0600	130.0863	65.5468			1

Monoisotopic mass of neutral peptide Mr(calc): 1386.7606
Variable modifications:
: NA_INSERTION
Ions Score: 73 Expect: 0.0013
Matches : 9/114 fragment ions using 14 most intense peaks (help)

MASCOT : Nucleic Acid Databases MATRIX SCIENCE

We have a nice example of a possible base insertion.