



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.

NA Translation	
1. The Standard Code (transl_table=1)	
By default all transl_table in GenBank flatfiles are equal to id 1, and this is not shown. When transl_table	
AAs = FFLLSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRVVVVAAAADDEEGGGG Starts =	
AAs = FFLLSSSSYY**CCWWLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG Starts =	
3. The Yeast Mitochondrial Code (transl_table=3) AAs = FFLLSSSSYY**CCWWITTTPPPPHHQQRRRRIIMMITTINNKKSSRRVVVVAAAADDEEGGGG Starts =	
MASCOT : Nucleic Acid Databases © 2012 Matrix Science	{MATRIX \ {SCIENCE}

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.

Inc con Inco 19	vorites <u>T</u> ools <u>H</u> elp							
🔶 Favorites 🏾 🏉 Mas	cot Database Maintenance		- 🖞 🗸	5	🖶 🔻 Page	<u>Safety</u>	T <u>o</u> ols ▼	• •
Database Manager Databases (19) Parse rules (17) Tasks (0) Settings New database Enable predefined definition Create new Synchronise custo definitions	Custom defini Database name: EST_Example Database type (?) Amino acid (protein databa Nucleic acid (DNA database (Note that database type can Sequence directory (?) Cingentumascoliseourgene	nnot be changed later.)						
	Database files will be located created if it does not already	in the directory EST_Exar exist.	mple in the seque	ence directo	ry. The nev	directory	will be	
	 Database files will be located created if it does not already Download database files FASTA file URL 	in the directory EST_Example exist.	<i>mple</i> in the seque	ence directo	ry. The nev	directory	will be	
	 Database files will be located created if it does not already Download database files FASTA file URL 	in the directory EST_Exar exist. (?)	<i>nple</i> in the seque	ence directo	ry. The nev	/ directory	will be	
	Database files will be located created if it does not already Download database files FASTA file URL Version file URL (optional	in the directory EST_Exar exist. (?)	<i>mple</i> in the sequ	ence directo	ry. The nev	v directory	will be	
	Database files will be located created if it does not already Download database files FASTA file URL Version file URL (optional	in the directory EST_Exar exist. : (?) al) (?)	<i>mple</i> in the sequ	ence directo	ry. The nev	v directory	will be	
	Database files will be located created if it does not already Download database files FASTA file URL Version file URL (optiona Reference file URL (optiona	in the directory EST_Exar exist. : (?) al) (?)	<i>nple</i> in the sequ	ence directo	ry. The nev	v directory	will be	
	Database files will be located created if it does not already Download database files FASTA file URL Version file URL (optiona Reference file URL (option	in the directory EST_Exar exist. : (?) al) (?) pnal) (?)	<i>mple</i> in the sequ	ence directo	ry. The nev	v directory	will be	
	Database files will be located created if it does not already Download database files FASTA file URL Version file URL (optiona Reference file URL (optio	in the directory EST_Exar exist. : (?) al) (?) onal) (?)	<i>nple</i> in the sequ	ence directo	ry. The nev	v directory	will be	
	Database files will be located created if it does not already Download database files FASTA file URL Version file URL (optional Reference file URL (optional Optional) I will copy the files to the	in the directory EST_Exar exist. (?) al) (?) e database directory (?)	<i>nple</i> in the sequ	ence directo	ry. The new	v directory	r will be	
	Database files will be located created if it does not already Download database files FASTA file URL Version file URL (optiona Reference file URL (option I will copy the files to the If you have chosen to downloa' You can follow the progress in downloaded.	in the directory EST_Exar exist. ; (?) al) (?) e database directory (?) ad files from a remote sen the task list. The databa	nple in the sequi ver, the downloa se configuration	ence directo ads will be sc can be com	ry. The new	a backgrout the files the	und task	c. en

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.

Contraction of the second	
C Matrix Science	e - Database Maintenance - Windows Internet Explorer
S - 4	http://yabby/mascot 🔻 🗟 😝 🗙 👂 Bing 🖉 👻
Eile Edit	(iew Favorites Iools Help
🚖 Favorites	🍘 Matrix Science - Databa 🛛 👘 🔻 🖾 🔻 🖃 🖷 💌 Bage 💌 Safety 🕶 🧇
Mascot	Database Maintenance: Edit Database Definitions
Help Wir	dow
	· · · · · · · · · · · · · · · · · · ·
Select	EST_human -
Name	EST_human Active Inactive
Path	C:/inetpub/mascot/sequence/EST_human/current/EST_human_".f
	Threads 24 Local ref file
Taxonom	v FST human FASTA with TaxID v
Rule to p	arse accession string from Fasta file:
	Rule 6: ">\(gi [0-9]*\)"
Rule to p	arse description string from Fasta file:
	Rule 7: ">[^]* \(.*\)"
Rule to p	arse accession string from local reference file:
	no local reference file •
Source a	nd parse rule for full text report (optional):
Host	eutils.ncbi.nlm.nih.gov Port 80
Path	/entrez/eutils/efetch.fcgi?rettype=gb&retmode=text&db=nucleotic
	Rule 11: "\(LOCUS .*\)"
	Test this definition Delete this definition
	New definition Edit parse rules Edit options
No	changes are written to marcet dat until you choose: APPLY
NO	Changes are written to mascouldt and you choose.
Done	🕵 Local intranet Protected Mode: Off 🍕 🔻 🖏 100% 👻
Nucle	MATRIX

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

Select Summary Report (iPRG200	8 UniPr +	ام	∞ - ↑
Most Visited Getting Started	💊 Latest Headlines 🔮 Google 🥃 Overview (Java 2 Platf 🐡 The World Clock - TL 💽 Overview (Java Platfor 🗋 Matrix Science - Home 🗋 Matrix Sci	ence - Datab	
(SCIENCE) IVIASCO	Search Results		
User :	PAE		
Email :			
Search title :	iPRG2008 UniProt Mouse		
MS data file :	D:\iPRG2008\mgf\merged.mgf		
Database :	UniProt_mouse mouse_20120418 (55190 sequences; 26879425 residues)		
Timestamp :	30 Apr 2012 at 10:33:41 GMT		
Enzyme :	Trypsin/P		
Fixed modifications :	<pre>iTRAQ4plex (K), iTRAQ4plex (N-term), Methylthio (C)</pre>		
Variable modifications :	Acetyl (Protein N-term), Gin->pyro-Glu (N-term Q), Oxidation (M)		
Mass values :	Monoisotopic		
Protein Mass :	Unrestricted		
Peptide Mass Tolerance :			
Fragment Mass Tolerance:	± 0.6 Da		
Max Missed Cleavages :	1 Per entre		
instrument type :	ESI-TRAP		
Protein hits :	F20022 78 KDa glucose-regulated protein 05-Mus musculus 0N-KBpa5 F2-1 SV-3 028850 Cytochrome P-5, incorm CRA, 06-Mus musculus 0N-Cyto F2-3 SV-1 028850 Cytochrome P450 2C29 C3-Mus musculus 0N-Cyto F2-4 020850 Cytochrome P450 2C3 020850 Cytochrome P450 2C3 020850 Cytochrome P450 2D3 020850 Cytochrome P450 2D4 020760 Uncharacteried protein O3-Mus musculus 0N-Cytol F8-4 SV-1 020760 Uncharacteried protein D3-Mus musculus 0N-Cytol F8-1 SV-2 020760 Uncharacteried protein D3-Mus musculus 0N-Cytol F8-1 SV-2 020761 Cytochrome P450 2A12 03-Mus musculus 0N-Cytol F8-1 SV-2 020762 Cytochrome P450 A013 03-Mus musculus 0N-Cytol F8-1 SV-2 020762 Cytochrome P450 A011 03-Mus musculus 0N-Cytol F8-3 SV-1 020762 Cytochrome P450 A010 04-Mus musculus 0N-Cytol F8-3 SV-1 020763 Cothrom P450 A010 04-Mus musculus 0N-Cytol F8-3 SV-1 020764 Long-chain-fatty-acid-c-CA 1igase 1 06-Mus musculus 0N-Kytol 1F8-1 SV-1 020765		,
	· Nuclaic Acid Databasas	<i>∫MA</i> 2	TRE

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.

Beer : PAE Email : : Email	Select Summary Report (IPRG200 File Edit View History Bookma Select Summary Report (IPRG20 Select Summary Report (IPRG20 Most Visited Selecting Started (MATRIX) Mascon	EST_moure_I - Macilla Fredor fris Tools: Edio % EST + m_muncel_24_06.4f cg/master_results plffiles_%2Fdats%2F001432.4ar8_queryfilts=sl66EPTYFEs select8_sigt C C Google p CC CC C CC C
	User Email MS data file Database Timestamp Enzyme Fixed modifications Variable modifications Variable modifications Protein Mass Teptide Mass Tolerance Fragment Mass Tolerance Max Missed Cleavages Instrument type Number of queries Protein hits	PAE 1950/2008 EST_House D:11950/2008/BGT_House D:11950/2008/BGT_House D:11950/2008/BGT_House D:11950/2008/BGT_House D:11950/2008/BGT_HOUSE D:11950/2008/BGT_HOUSE 100/2008/BGT 100/2008/BGT 100/2008/BGT 100/2008/BGT 100/2008/BGT 100/2008/BGT 100/2008/BGT 100/2008/BGT 100/2008/BGT 100/2008/B

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.

Significance Threshold p> 0.05 Max. number of hds AUT0 Show Percolator scores Standard scoring © MudPTI scoring @ Lons score or expect cut-off 0.5 Show sub-sets 0 Show sub-sets 0 Show sub-sets 0 Show sub-sets 0 Show sub-sets 0 Reverse 1 #All queries Unassigned © Bdow homology threshold Image: Statistic 0 Max musclus cDNA clone HAGE 1303045 5' similar to gb:N22465 CTTOCHROME B5 (HINGAN);, mBNA sequence Operation 0 Descrete Mir (weight Mr clone) Delta Mines Score Expect Enab. Molique Peptide 15020 733.0084 1474.0022 1473.7399 0.2064 0 74 0.0018 1 K.YTTIBEIGN.H 18018 18021 18023 18023 18023 20552 787.04662 1572.0776 1571.8065 0.2712 0 76 0.0011 1 K.YTTIBEIGN.H 18018 18021 18023 18023 18023 20552 787.04662 1572.0776 1571.8065 0.2712 0 76 0.001 1 K.YTTIBEIGN.H 18018 18021 18023 18023 18023 20552 787.04662 1572.0776 1571.8065 0.2712 0 76 0 2 0.023 1 K.YTTIBEIGN.H 18018 18021 18023 18023 1923 0123 20552 784.1341 655.483 1655.483 7 0.2047 0 .0247 0 62 0.023 1 K.YTTIBEIGN.H 18018 18021 18023 18023 1923 0123 0123 0123 0123 0123 0123 0123 01		st visited 👹 Getting started 🚵 Latest Headlines 🏹 Google 💆 Overview (Java 2 Platt 💎 The v	World Clock – Ti Overview (Java Platfor Matrix Science - Home Matrix Science - Datab
Standard scoring © MudPT scoring ® [ons score or expect cut off 0.5 Show mab-sets 0 Show pop-ups © Suppress pop-ups © Requere bod red ResSearch @ All queries © Unassigned © Below homology threshold © Below identity threshold Image: Standard Score: 631 Matches: 19(15) Sequences: 4(4) emBAI: 1.96 Frame: 1 mb24602:15 Stares mouse pSMMP15.6 Num susculus colls clone DAGC:330343 5' similar to gb:H22865 CTIOCHROME B5 (HIMAN);, mBNA sequence Query Observed Mr(expt) Mr(calc) Delta Miss Score Expect Rank Onique Peptide 12022 78.0064 1474.0022 1473.7959 0.2064 0 7 4 0.0011 1 K.TYTIOEHIPOR.8.2051 20555 797.0462 1572.0778 1571.0055 0.4201 0 7 4 0.0011 1 K.TYTIOEHIPOR.8.2051 20552 20.061 1572.2066 1571.0055 0.4201 0 7 6 0.0011 1 K.TYTIOEHIPOR.8.2051 20553 20.061 1572.2066 1571.0021 473.7978 0 104 6.4e=07 1 R.EQACODATESHFEWORSTAR.8.2051 20515 784.1341 625.0483 1654.6437 0.2027 0 .33078 0 104 6.4e=07 1 R.EQACODATESHFEWORSTAR.8.20412 30412 30412 30428 3		Significance threshold p< 0.05 Max. number of hits AUTO	Show Percolator scores
Show pop-ups Suppress pop-ups Require bold red Review * All queries Unassigned Bedow identity threshold I: gli228557 Mass: 18637 Socre: 631 Matches: 19(13) Sequences: 4(4) emBAI: 1.96 Frame: 1 mb24f02:11 Socre: 631 Matches: 19(13) Sequences: 4(4) emBAI: 1.96 Frame: 1 mb24f02:11 Socre: 631 Matches: 19(13) Sequences: Expect Eank Onique Peptide Sequence: Expect Eank Onique Peptide 20552 797.0642 1572.0778 1571.0655 0.4031 74 0.0031 I K.YVTUERUFOR: # 15013 18023 15023 15022 20552 797.0642 1572.0778 1571.0655 0.4031 0 74 0.0031 I K.YVTUERUFOR: # 20513 15023 15023 15022 20553 797.06425 1572.0778 1534.6437 0.2047 0 0 0.023 I K.YVTUERUFOR: # 20513 22316 2073.5378 104 6.46-677 R RequirebookTista.E 20427 20423 20428 20428 20428 20428 20428 20428 20428 20428 20428 20		Standard scoring MudPIT scoring Ions score or expect cut-off 0.5	Show sub-sets 0
Referent • All querie: • Unassigned • Bedow thomology threshold • Bedow identify threshold • gi1228557 Mass: 18657 Score: 631 Matches: 19(13) Sequences: 4(4) em2Ai: 1.96 Frame: 1 mb24f02.11 Soares mouse p3NHT19.5 Nus musculus CDNA clone HAGE:330343 5' similar to gbt/22865 CTOCKROME B5 (HUMAN);, mRNA sequence (Geory Concerns) Mass: 18657 Score: 631 Matches: 19(13) Sequences: 4(4) em2Ai: 1.96 Frame: 1 mb24f02.11 Soares mouse p3NHT19.5 Nus musculus CDNA clone HAGE:330343 5' similar to gbt/22865 (HUMAN);, mRNA sequence (Geory Concerns) Mass: 1873 Soare: 10: 10 K.YTUERENDOR.8 20552 2055 205 2055 2055 205 205 2055 205		Show pop-ups 💿 Suppress pop-ups 💌	Require bold red
mb24f02.rl Soares mouse p3MR719.5 Mus musculus cDNA clone PAGE:330348 9' similar to gb:M22865 CTTOCHROME BS (HURAN);, mRNA sequence Qeerg Observed Mr(equt) Mr(clo1) DeltA Miss Score Expect Eank Unique Peptide 16020 738.0084 1474.0022 1473.7399 0.2064 0 74 0.0018 1 K.YTHIGELHEDDA.52051 16023 16023 20552 797.0462 1572.0786 1571.0055 0.2712 0 76 0.0011 1 K.YTHIGELHEDDA.52051 20552 797.0462 1572.0786 1571.0055 0.2712 0 76 0.0021 1 K.YTHIGELHEDDA.52051 20552 797.0462 1572.0483 1564.4647 0.2047 0 42 0.023 1 K.YTHIGELHEDDA.52051 20552 797.0483 1554.4647 0.2047 0 42 0.023 1 K.FTLERHEDDA.52051 20552 797.0483 1554.4647 0.2047 0 42 0.023 1 K.FTLERHEDDA.52051 20553 794.1341 2349.3805 2349.0227 0.3578 0 104 6.4e-67 1 R.EQAOGATEMTEDVOHSTDAA.E 30412 30425 30426 30428 30438 30408 40000000000000000000000000	Re-S	Search • All queries Unassigned Below homology threshold Below g1 1288567 Mass: 18637 Score: 631 Matches: 19(15) Sequence	videntity threshold ces: 4(4) emPAI: 1.96 Frame: 1
Construction Construction Construction Construction Construction 15021 798.0044 144.0022 147.198 0.2044 0 74 0.0011 1 K.TYTLBEIGK.H 18021 18022 18023 1		mb24f02.rl Soares mouse p3NMF19.5 Mus musculus cDNA clone IMAGE:33 Onery Observed Mr(empt) Mr(calc) Delta Miss Score Expect F	0363 5' similar to gb:M22865 CYTOCHROME B5 (HUMAN);, mRNA sequence Rank Unique Pentide
20552 097.0642 1572.0778 1571.0605 0.712 0 76 0.0011 1 K.TYIIOKINDOR.5 20554 205.0661 1572.2666 1572.0768 1572.0768 1572.077 1 K.TYIIOKINDOR.5 22335 201.501.061 1552.0661 1563.0643 1654.0437 0.2047 0 62 0.023 1 K.FVEIDRUMENDE.8 20511 22335 22335.0472 23402.027 0.3370 0 (100) 1.9e-06 1 R.EQACODATENTRZOVENTAR.E 30412 30412 30426 30426 30426 30428 30432 20113 177.6897 2845.0472 30402 104 6.4e-07 1 R.EQACODATENTRZOVENTAR.E 30412 30412 30412 30426 30426 30428 30428 30428 30428 30432 Proteins matching the same set of peptides: 1 REQACODATENTRZOVENTAR.E 30412 30412 30417 30428 30426 30428 30428 30428 30428 30432 g11159200 Mass: 18789 Score: 631 Matches: 19(15) Sequences: 4(4) Frame: 3 m35607.r1 Sore: 631 Matches: 19(15) Sequences: 4(1) Frame: 1 1 1 m12408.r1 Sore: 501 Matches: 19(15) Sequences: 4(1) Frame: 1		18020 738.0084 1474.0022 1473.7959 0.2064 0 74 0.0018	1 K.YYTLEEIOK.H 18018 18021 18023 18025
20553 052.0661 1572.2366 1572.2367 1572.2367 0.0400 0 (100) 1.9=-06 1 R.FUERENDORSTAR.E 30412 30412 30412 30428 304208 30428 30428 30428 30428 30428 30428 30428 30428 30428 30428 30		20552 787.0462 1572.0778 1571.8065 0.2712 0 76 0.0011	1 K.TYIIGELHPDDR.S
2236 622.5314 1655.0483 1655.0483 1654.0437 0.2047 0 62 0.023 1 R.PLEREMPOINSTDALE 3042 20113 175.0897 234.5047 0.3207 0.3207 0.100 1.se.0A0CDATENFEDVORSTDALE 30427 30113 784.1341 2349.3805 2349.0227 0.3378 0 104 6.4e-07 1 R.EQACODATENFEDVORSTDALE 30422 3		<u>20563</u> 525.0861 1572.2366 1571.8065 0.4301 0 (57) 0.077	1 K.TYIIGELHPDDR.S 20551
20113 1173.6897 2349.0227 0.3420 0 (100) 1.9=-06 1 R.BQACODATENTEDVORSTDAR.E 30127 20115 784.1341 2349.0227 0.3578 0 104 6.4e=07 1 R.BQACODATENTEDVORSTDAR.E 3012 30425 30426 30422 3042		<u>22386</u> 828.5314 1655.0483 1654.8437 0.2047 0 62 0.023	1 K. FLEEHPGGEEVLR. E
30115 784.1341 2349.0227 0.3578 0 104 6.4e-07 1 R.EQAGGAATENTERVENTERAR.E 30112 3012 3012 3012 3012 3012 3012 301		<u>30413</u> 1175.6897 2349.3647 2349.0227 0.3420 0 (100) 1.9e-06	1 R.EQAGGDATENFEDVGHSTDAR.E <u>30427</u>
		<u>a11366609</u> Mass: 18799 Socre: 631 Matches: 19(15) Sequent md5e07.1: Soares mouse embryo NMEI3.5: 14.5 Nus musculus cNN. clo. 113609.2: Mass: 20396 Socre: 631 Matches: 19(15) Sequent m15608.7: 13 Soares mouse embryo NMEI3.5: 14.5 Nus musculus cNN. clo. 211509570 Mass: 18741 Socre: 631 Matches: 19(15) Sequent m156471.7: 1 Soares mouse embryo NMEI3.5: 14.5 Nus musculus cNN. clo. 211509262 Mass: 2019 Socre: 631 Matches: 19(15) Sequent m56407.7: 1 Soares mouse embryo NMEI3.5: 14.5 Mus musculus cNN. clo. 211509270 Mass: 2025 Socre: 631 Matches: 19(15) Sequent m56407.7: 1 Soares mouse embryo NMEI3.5: 14.5 Mus musculus cNN. clo. 21150307 Mass: 2025 Socre: 631 Matches: 19(15) Sequent m56407.7: 1 Soares mouse embryo NMEI3.5: 14.5 Mus musculus cNN. clo. 21150307 Mass: 2025 Socre: 631 Matches: 19(15) Sequent m55407.7: 1 Soares mouse embryo NMEI3.5: 14.5 Mus musculus cNN. clo. 21150307 Mass: 2025 Socre: 631 Matches: 19(15) Sequent m65506.7: 1 Soares mouse embryo NMEI3.5: 14.5 Mus musculus cNN. clo. 21150307 Mass: 2025 Socre: 631 Matches: 19(15) Sequent m65506.7: 1 Soares mouse embryo NMEI3.5: 14.5 Mus musculus cNN. clo. 21150307 Mass: 2025 Socre: 631 Matches: 19(15) Sequent m65506.7: 1 Soares mouse embryo NMEI3.5: 14.5 Mus musculus cNN. clo. 21150307 Mass: 2025 Socre: 631 Matches: 19(15) Sequent m65506.7: 1 Soares mouse embryo NMEI3.5: 14.5 Mus musculus cNN. clo. 21150407 Mass: 2025 Socre: 631 Matches: 19(15) S	<pre>ces: 4(4) Frame: 1 ne HAGE:1375405 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 1 ne HAGE:14565 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 2 ne HAGE:445398 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 1 ne HAGE:45585 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 1 ne HAGE:45595 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 1 ne HAGE:453391 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 1 ne HAGE:453391 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 1 ne HAGE:453391 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 1 ne HAGE:453391 5' similar to gb:H22865 CYTOCHROME B5 (HUMAN);, mRNA sequence ces: 4(4) Frame: 1 </pre>

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.

Select Summary Report (IPRG200 File Edit View History Bookma Select Summary Report (IPRG200	8 EST_mouse) - Mozill irks <u>T</u> ools <u>H</u> elp 38 EST +	a Firefox		
S S	ng/mascat.24.0.64/c	girmater realts pMMes. NGF4ar%1700104000047 Google © Overview Usea 2 Plaff. ☆ The W Results monae griparged.mgf griparged.mgf griparged.grip j.jTAQ40plex (H-term), Methylthio (C in H-term), Oln->pyre-Olu (H-term Q) mk24603.rl Soares mouse p3NH15.5 M mg16c12.rl Soares mouse p3NH15.5 m42007.rl Soares mouse p3NH15.5 m42005.rl Soares mouse p3NH15.5 massed part of the result of the resul	001412.det#, queryiets alf鎐FsteleteR.get ? ♥ ♥ ord#Clock - TL ♥ Overview (Java Platfer. Matrix Scie 000529140 residues) 1 0 0 0 0 0 0 0 0 0 0 0 0 0</th><th>Coopie Score Score Score Score Score Score Score Scor</th></tr><tr><td></td><td>· Nuo</td><td>laia Aaid Databaa</td><td></td><td>(MATRIX</td></tr></tbody></table>	

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.

Home - UniGene - NCBI	+ v.ncbi.nlm. nih.gov /unigene			☆ ⊽ Ć 🛃 = Goo	nle	م م	.
Most Visited 🚱 Getting Starte	d 🔊 Latest Headlines 💐 Google	Overview (Java 2 Platf 🐣 The Worl	d Clock - Ti 🧿 Overview (la	va Platfor Matrix Sci	ence - Home 🗍 Matr	ix Science - Da	atab
S NCBI Resources ⊡ H	łow To 🕑		_		N	ly NCBI Sigr	n In
UniGene	-						
official off	Limits Adv	anced			Search	н	elp
ACGT	GT	UniGene					
CGTCGC	TAC GT						
CGTCGC	TAC	UniGene computationally identified	es transcripts from the sam	e locus; analyzes expre	ession by tissue, ag	e, and	
CGT	CT.	Treatin status, and reports related	u proteins (protesi) and ci	one resources.			
C .	TAC						
Using UniGene		UniGene Tools		Also of interest			
Getting Started		UniGene Statistics		BioSample			
TAQ		Library browser		EST sequences			
Query Tips		Digital Differential Display (DDD)		Gene			
ProtEST		Downloads/FTP		HomoloGene			
DDD				Map Viewer			
Clustering by transcripts				Short Read Archive (SRA	ŋ		
Clustering by genomes							
You are here: NCBI > Genes & Expr	ession > UniGene				W	ite to the Help D	Desk
GETTING STARTED	RESOURCES	POPULAR	FEATURED		NCBI INFORMATION		
NCBI Education	Chemicals & Bioassays	PubMed	Genetic Testing R	legistry	About NCBI		
NCBI Help Manual	Data & Software	Nucleotide	PubMed Health		Research at NCBI		
							_

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.

Select Summary Report (IPRG2D) File Edit View History Bookm Select Summary Report (IPRG2D & S S B bogs	88 BST_mouse) - Mazilla Firefox arks_Tonk: Holp 08 BST_m + mg/mascet_2.4.9_54/cgi/master_results.p1fflee%2Fdata%2F30130430%2FF001	1432.det&_queryfist=sil&SEPTYPE=select&_sigt $\stackrel{\sim}{\cong} e \ C$	P • *
Most Vaided Cetting Started Sciences Masses Sciences Masses Sciences Masses Search title Stata file Database Timestamp Ensyme Transfer Sciences Protein Mass Tolerance Ass Vaides Tolerance Frequent Mass Tolerance Ass Misses Cleavages Frotein hits	Latest Headliner * Google C Overview (Java 2 Plant * The World Scance D Results PA8 PA8	d Clock - TL Overview (Java Platfor Matrix Science - Home Matrix 5 529140 residues) stidation Max. number on More Max. Number on	<pre>Kience - Datab 5 CYTOCHROME B5 (HIDANI); to gbiN2265 CYTOCHROME 5', mRNA sequence r to SN:R14A_YEAT P361C ar to gbiN61855 CYTOCHROME P450 III2 g-cds1 Mouse surfeit loc A0312A06 IMAGE:30734008 s CDNA clone mia01045 5' 053 CYTOCHROME P450 IA2 D profein_5_(d)mouse_ref</pre>
ASCOT	: Nucleic Acid Database	S © 2012 Matrix Science	(MATRIX SCIENCE

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

S Composition Source 1, 24,947-07 matter, entropy 1700,132,2485, query first addREPTYPEs selectADUE (" " C)" Cogle // C ogle // C og	★ 5 € Control Cont	geng march_14.0.04/cg/marker_result.gfMes.XF20113410452/F001422.detB_quergidmt.aHBAEPTYPErsetectBAURE ☆ * C	
<pre>(*) Most Windel & Gating Sander & Latest Headines & Google © Overview Uses 2 Mart. ** The World Clock - T. © Overview Uses Platfor. ** Matrix Science - Datab. Tratement is 251-TRAF France of queries : \$3131 Frotein hits : \$3131 Frotein</pre>	Motivade & Geting Sur Tastrument - Tastrument - Tastru	Construct Headines of Google Coverview (Java 2Platf., ** The World Clock-TL. Coverview (Java Platfor., Matrix Science - Home Matrix Science - Databa FSI-TRAF. 1 Salist 1 Salist	
Tastrument type i BSI-TRAP Frotein hits i Bs.3016 Frotein hits i Bs.301 Frotein hits	Instrument type Number of queries Protein hits	: BS1-TRAP : S3191 : Mm.31018 CybS Cytochrome b-5 Mm.31019 Mapa5 Heat shock protein 5 Mm.15796 Mgs1 Nicrosomal glutashione 5-transferame 1 Mm.15537 Cypi2d Cytochrome P450, family 1, subfamily a, polypeptide 2 gl128843848 AGENEOUKT_12238258 NIH MGC 136 Mum musculus cDNA clone HAGE:30289533 5', mRNA sequence Mm.20766 Cypi2d Sytochrome P450, family 2, subfamily a, polypeptide 2 Mm.16660 P4hb Frolly 4-hydroxylase, beta polypeptide Mm.6660 Rdh Fatinol dehydrogenae 7 Mm.2083011 Rp17m Ribosomal protein 17A Mm.2083012 Cypial Cytochrome P460, family 3, subfamily a, polypeptide 11 Mm.209501 Cesif Carboxylestrame 17 Mm.209501 Cesif Carboxylestrame 13A Mm.209502 Cesid Carboxylestrame 10 Mm.20910 Cesif Carboxylestrame 10 Mm.20910 Had17b6 Hydroxytertoid (17-beta) dehydrogenae 6	
Number of queries : 33391 Protein hits : 10.0128 (yp6 Cytochrome b-5 Bh.300120 (yp6 Cytochrome b-1 Bh.300120 (yp6 Cytochrome b-1 Bh.300120 (yp6 Cytochrome b-1 Bh.300120 (yp6 Cytochrome b-1 Bh.300120 (yp6)Cytochrome b-1 Sh.12892.59 Mil MGC.196 Mas maculus CNR Alone HAGE13028553 5', mRNR sequence Ha.1660 RdR Teclol Chytopromes 7, subfamily e, polymeptide 2 Ha.1660 RdR Teclol Chytopromes 7, subfamily e, polymeptide 11 Ha.2010 Cecif Carboy/setresse 2R Ha.2010 Cecif Carboy/setresse 3R Ha.2010 Cecif Carboy/setresse 3R Ha.2010 Cytochrome P450, family 7, subfamily e, polymeptide 11 Ha.2010 Cecif Carboy/setresse 3R Ha.2010 Cecif Carboy/setresse 3R Ha.2010 Cytochrome P450, family 7, subfamily e, polymeptide 12 Ha.2010 Cytochrome P450, family 7, subfamily e, polymeptide 12 Ha.2013 Cytochrome P450, family 7, subfamily e, polymeptide 12 Ha.2013 Cytochrome P450, family 7, subfamily e, polymeptide 12 Ha.2013 Photoemal protein, large, P0 Ha.2013 Philosemal protein, large, P0 Ha.2017 Philosemal protein, large, P0 Ha.2017 Philosemal protein, large, P0 Ha.2017 Philosemal protein, large, P1 Ha.2017 Philosemal protein, large, P1 Ha.2017 Philosemal protein, large, P2 Ha.2017 Philosemal protein, large, P1 Ha.2017 Philosemal protein, large, P2 Ha.2017 Philosemal protein, large, P1 Ha.2017 Philosemal protein, large, P2 Ha.2017 Philosemal protein, large, P3 Ha.2017 Philosemal protein, large, P2 Ha.2017 Philosemal protein, large P2 H	Number of queries Protein hits	: 33191 : 33191 : Ma.1012 Mar.1025 Mar.1035160 Mar.10350 Mar.10350 Mar.10350 Mar.10350 Mar.10450 Mar	
Protein hits (pp Cycle Cyclechows b-5 He. 300100 Hpps Heat Shock protein 5 He. 10797 Mpst: Microsomal glutamined y 1. subfamily 4, polypeptide 2 112995464 AURCOMERT 1222055 WIM MGC 154 New mencular Coll Coll 202555 5', mBNR sequence He.20764 Cyp2c3 Cyclechows P450, family 2, subfamily 6, polypeptide 29 He. 16605 PH br Polyl 4-hydroxylase, beta polypeptide He. 6656 Cycle Polyl 4-hydroxylase, beta polypeptide He. 2010 Cesif Carboxylererase 17 He. 2010 Cesif Carboxylererase 10 He. 2010 Cycle Corboxe P450, family 2, subfamily 4, polypeptide 12 He. 174372 Cycle Cycle Poly 14-hydroxylere 2 family 4, polypeptide 12 He. 174372 Cycle Cycle Poly 14-hydroxylererase 1 family 4, polypeptide 10 He. 2010 Dy Columoroyltansferase 2 family 4, polypeptide 10 He. 2010 Dy Columoroyltansferase 2 family 4, polypeptide 10 He. 2010 Dy Columoroyltansferase 2 family 4, polypeptide 10 He. 2010 Dy Riboxemal protein, large F2 He. 37138 Phibosemal protein, large F2 He. 37138 CycleStrox Portein, St. Max merculus CMR clone HAGE:65107 5', mBNR sequence He. 44134 Cycle Frocein Mild is informer associated 4 He. 379379 CycleSt Cyclochrome F450, family 2, subfamily c, polypeptide 54 Gli 122044609 ADECOURT 8001020 Nit Mid (2) AM Me marculus CMR clone HAGE:6510216 5', mBNR sequence He. 44134 Cycle Frocein Mild (19007041 Nit marculus CMR clone HAGE:6510216 5', mBNR sequence He. 45137 Nithel Histon Columbia Family mild is more associated 4 He. 379579 CycleSt Cyclochrome F450, family 2, su	Frotein hits	: Mm.31010 CybS Cytochrome b-5 Mm.31010 Kaps5 Nest shock protein 5 Mm.14796 Mgs1 Nicrosemal qlutachicos 5-transferase 1 Mm.15537 Cypl24 Cytochrome P450, family 1, subfamily a, polypeptide 2 Gui2084564 AGENCOURT_12238135 NIT NGC 136 Num musculus CDNA clone INGAE:3028533 S', mRNA sequence Mm.20764 Cypl24 Cytochrome P450, family 2, subfamily c, polypeptide 29 Mm.15667 P41b Frolly 4-hydroxylase, beta polypeptide Mm.302864 Cypl241 Cytochrome P450, family 3, subfamily a, polypeptide 11 Mm.302864 Cypl241 Cytochrome P450, family 3, subfamily a, polypeptide 11 Mm.302864 Cypl241 Cytochrome P450, family 3, subfamily a, polypeptide 11 Mm.302864 Cypl241 Cytochrome P450, family 3, subfamily a, polypeptide 11 Mm.302864 Cypl241 Cytochrome P450, family 3, subfamily a, polypeptide 11 Mm.302864 Cypl241 Cytochrome P450, family 3, subfamily a, polypeptide 11 Mm.302864 Cypl241 Cytochrome P450, family 3, subfamily a, polypeptide 11 Mm.302864 Cypl241 Cytochrome P450, family 3, subfamily a, polypeptide 11 Mm.302865 Cesid Cathoxylestrase 17 Mm.202903 Cesid Cathoxylestrase 18 Mm.202105 Msd17b6 Hydroxylestrase 18 Mm.202105 Msd17b6 Hydroxysteroid (17-beta) dehydrogenase 6 Mm.2017b1 Mydroxyteroid (17-beta) dehydrogenase 6	
<pre>min.2016 im.2017</pre>		<pre>Important provides a second protection for a second s</pre>	
<pre> Topis Cycochema #440. family 1. subfamily 6. polymptide 2 gli2945444 ANDCOMER_1223259 Wilk MGC 154 Num smeculae CANA close HAGAL SIG28553 5', mENR sequence Ma.2074 Cypc23 Cytochcome P450, family 2, subfamily c, polymptide 29 Main 6656 Fibb Foolyl 4-hydroxylase, beta polymptide Ma.6657 Kair Retinol dehydrogenase 7 Ma.32046 CypSall Cytochcome P450, family 3, subfamily a, polymptide 11 Ma.530266 CypSall Cytochcome P450, family 3, subfamily a, polymptide 11 Ma.530266 CypSall Cytochcome P450, family 3, subfamily a, polymptide 11 Ma.530266 Cascowylestcrase 17 Ma.53027 CypIII Cytochcome P450, family 2, subfamily a, polymptide 12 Ma.530261 Tesacribed locax, subfamily a, polymptide 12 Ma.530264 Main Catabaylestcrase 10 Ma.2010 CypIII Cytochcome P450, family 2, subfamily a, polymptide 12 Main 1945 Main Cytochcome P450, family 2, subfamily a, polymptide 12 Main 2010 Main 2010</pre>		 Instant Constant Parks (17-beam parks) Instant Parks) Instant Parks (17-beam parks) Instant Parks) Instant Parks (17-beam parks) Instant Parks) Ins	
<pre>ili2004.cond ili2004.cond</pre>		<pre>s1129854649 AdEEcoURT 12238259 1111 MOC_134 Kum musculus cDNA clone IMAGE:00289533 5', mRNA sequence Ma.20764 Cyp2c3 Cytochrome P450, fmaily 2, subfamily c, polypeptide 29 Ma.6666 P4hb Froly1 4-hydroxylase, beta polypeptide Ma.6393 AdF Retiral dehydrogenase 7 Ma.393841 Cyp3all Cytochrome P450, family 3, subfamily a, polypeptide 11 Ma.3294001 Transcribed locus, strongly similar to NP_001004379.1 605 ribosomal protein L7a [Gallus gallus] Ma.2010 Cost Controlyceterine 17 Ma.393803 Cesid Catboxylestresse 10 Ma.20105 Cesid Catboxylestresse 11 Ma.20105 Cesid Catboxylestresse 11 Ma.20105 Cesid Catboxylestresse 15 Ma.20176 Myddroxysteroid (17-beta) dehydrogenase 6</pre>	
<pre></pre>		 18:.07744 Cyp2c39 Cytochrome P450, family 2, subfamily c, polypeptide 29 18:.16660 P4hb Prolyl 4-hydroxylase, beta polypeptide 29 18:.398371 Rylra Riboconal protein 1/7 18:.398371 Cytochrome P450, family 3, subfamily a, polypeptide 11 18:.328464 Cyp3all Cytochrome P450, family 3, subfamily a, polypeptide 11 18:.328461 Cytochrome P450, family 3, subfamily 600 ribosomal protein 1/7a 18:.328461 Cytochrome P450, family 5, subfamily 600 ribosomal protein 1/7a [Gallus gallus] 18:.328461 Cesia Cathoxylesterase 17 18:.22803 Cesia Cathoxylesterase 18 18:.22803 Cesia Cathoxylesterase 18 18:.20176 Hydroxysteroid (17-beta) dehydrogenase 6 	
Hein 16660 Fibb Frolyl 4-hydroxylase, beta polypeptide Hein669 RdN7 Retiol Gehydrogenase 7 Hein303071 Rp178 Ribosemal protein 17A Hein30214 Cypebrem 4960, fmaily 3, subfamily a, polypeptide 11 Hein30214 Cypebrem 4960, fmaily 3, subfamily a, polypeptide 11 Hein30215 Cesld Carboxylesterase 1A Hein20210 Cypebrem 4960, fmaily 7, subfamily a, polypeptide 12 Hein30210 Cypebrem 4960, fmaily 7, subfamily a, polypeptide 12 Hein30210 Cypebrem 4960, fmaily 7, subfamily a, polypeptide 12 Hein30210 Cypebrem 4960, fmaily 7, subfamily a, polypeptide 10 Hein30210 Cypebrem 4960, fmaily 7, subfamily a, polypeptide 10 Hein30210 Cypebrem 4960, fmaily 7, subfamily a, polypeptide 10 Hein30210 Cypebrem 4960, fmaily 7, subfamily a, polypeptide 10 Hein30215 Rp1p0 Riboremal protein, large F2 Hein30215 Rp1p0 Riboremal protein, large F2 Hein30215 Rp1p0 Riboremal protein, large F2 Hein30216 Cypebrem 4070 Hein30, hein30 Hein3		 Be.16660 Fibb Froly: 4-hydroxylase, beta polypeptide Ba.6690 AMP Retinol dehydrogenase 7 Be.393841 GypSall Cycohomse Fi50, family 3, subfamily a, polypeptide 11 Be.393801 Transcribed locus, strongly similar to NP_001004379.1 605 ribosomal protein L7a [Gallus gallus] Be.393803 Cesid Carboxylestriase 10 Be.393803 Cesid Carboxylestriase 10 Be.303804 Cesid Carboxylestriase 10 Be.303805 Cesid Carboxylestriase 10 Be.304106 Jint (J7-beta) dehydrogenase 6 	
<pre>Ref. Section 2. S</pre>		Ba. 6696 Rdh7 RetinOl dehydrogenase 7 Ma. 398971 Rp17a Ribocanal protein 1.72 3, subfamily a, polypeptide 11 Ma. 328244 CypSall Cyrochrome P450, family 3, subfamily a, polypeptide 11 Ma. 328461 CyroSall Cours, stronghy similar to NP_001004379.1 605 ribosomal protein L7a [Gallus gallus] Ma. 29210 Cesif Carboxylesterase 1F Ma. 292803 Cesid Carboxylesterase 1D Ma. 20210 Cesif Carboxylesterase 1D Ma. 20210 Cesif Carboxylesterase 1D Ma. 20213 Main Section 1(7-beta) dehydrogenase 6	
<pre>km.398271 April Cytochrome P450, fmaily 3, mbfamily a, polypeptide 11 km.328461 Cytochrome P450, fmaily 3, mbfamily a, polypeptide 11 km.328461 Transcribed locus, strongly similar to MP_001004375.1 603 ribosomal protein L7a [Gallus gallus] tm.228514 Crs5A Cathodysterturas 1A tm.22852 Crs5A Cathodysterturas 1A tm.22854 Crs5A Cathodysterturas 2A tm.22855 tm.22855 Crs5A Cathodysterturas 2A t</pre>		Ma.399871 Rp17a Rtboomal protein 17A Ma.3329401 Cyp3al Cycohorme P450, family 3, subfamily a, polypeptide 11 Ma.294001 Transcribed locus, strongly similar to NP_001004379.1 60S ribosomal protein L7a [Gallus gallus] Ma.29100 CesiG Cathoxylestronse 11 Ma.292003 CesiG Cathoxylestronse 10 Ma.292003 CesiG Cathoxylestronse 10 Ma.292003 CesiG Cathoxylestronse 10 Ma.292003 CesiG Cathoxylestronse 10 Ma.292003 CesiG Cathoxylestronse 10 Ma.29704 CesiG Cathoxylestronse 10 Ma.29718 CesiG Cathoxylestronse 10 Ma.29718 Mal7De Mydroxyteronid (17-beta) dehydrogenase 6	
<pre>https://www.intercommerce.projection.projection.pro</pre>		Mm.3232444 CypSall Cyrochrome P450, family 3, subtamily a, polymeptide 11 Mm.328641 Transcribed locus, strongly similar to NP_00100439.1 605 ribosomal protein L7a [Gallus gallus] Mm.29310 Cesif Carboxylesterase 17 Mm.292303 Cesid Carboxylesterase 3A Mm.292903 Cesid Carboxylesterase 10 Mm.20713 Mail/D6 Bydroxylesterase 10	
<pre>dm_200001 Testscience locus, strongry samilar to NP_001004979.1 605 Tibosomal protein 1/a [ualius gallus] tm_20101 cost dambayisercasa la tm_202010 cost dambayisercasa la tm_202020 Cesid Carboxylestcasa la tm_202020 cost dambayisercasa la tm_202020 cost dambayisercasa la tm_202020 cost dambayisercasa la tm_202020 cost cost cost cost cost cost cost cost</pre>		Ma.271001 Transoribed locus, strongly similar to Nr_001004379.1 Gos Ticosomal protein L/a [Uallus gallus] Ma.22100 Cost Control/Materians I Ma.292003 Cost Cathoxylestrons I Ma.292003 Cost Cathoxylestrons I Ma.29210 Cost Cathoxylestrons I Ma.2017 Mathra Mathra Mythoxysteroid (17-beta) dehydrogenase 6	
<pre> Const Lincody Sectors i Const Lincody Sectors i Const Lincody Sectors i Sector</pre>		Int.22110 Cesia Carboxylesteraes in Mm.229203 Cesia Carboxylesteraes 3A Mm.229203 Cesia Carboxylesteraes 1D Mm.26719 Had7b6 Hydroxytercial (17-beta) dehydrogenase 6	
<pre>ima_22200 Celd Catcopylercease 1D ima_22201 Middle Mydrogenase 6 ima_2271 Middle Mydrogenase 6(imagenase 6(imagenase 6) imagenase for the State of the State</pre>		Mm.292803 Cesid Carboxylesterase 1D Mm.26719 Hsd17b6 Hydroxysteroid (17-beta) dehydrogenase 6	
<pre>ima_2021s imailed Nytroxysteroid (17-bets) dehydrogenase 6 ima_20770 Cyp2d10 Cytochome 9450, family 2, subfamily 4, polypeptide 12 ima_174372 Cyp2d10 Cytochome 9450, family 2, subfamily 4, polypeptide 10 ima_20241 Cytochome 9450, family 2, subfamily 4, polypeptide 11 ima_20241 Cytochome 9450, family 2, subfamily 4, polypeptide 11 ima_20241 Cytochome 9450, family 2, subfamily 4, polypeptide 11 ima_20241 Cytochome 9450, family 2, subfamily 4, polypeptide 12 ima_20241 Cytochome 9450, family 2, subfamily 4, polypeptide 11 ima_202441 Cytochome 9450, family 2, subfamily 4, polypeptide 11 ima_202442 Pitat Forcian introduction 11 ima_202442 Pitat Forcian introduction 11 ima_202442 Pitat Forcian introduction 12, subfamily 4, polypeptide 54 gi12371440 Pitate Forcian introduction 2, subfamily 4, polypeptide 54 gi12371440 Restard Forcian introduction 1, subfamily 6, polypeptide 54 gi12371440 Restard Extra 2, subfamily 6, polypeptide 54 gi12371440 Restard Extra 2, subfamily 6, polypeptide 54 gi12371440 Restard Extra 2, subfamily 6, polymether 1 gi12371440 Restard Extra 2, subfamily 6, polymether 1 gi1237140 Restard Extra 2, subfamily 6, polymether 1 gi123714 Code synthetase 10ng-chain family member 1 gi123718 Apple Stochomone 113 pseudogene gi123718 Code Stochomone 113 pseudogene gi123718 Code Stochomone 113 pseudogene gi123718 Code Stochomone 2, family 2, subfamily 6, software 2, software 3, software 2, software 2, software 2, software 3, software 2, software 3, software 2, software 3, software 2, software 3, software 3, software 3, software 3, software 3, software 2, software 3, software 3,</pre>		Mm.26719 Hsd17b6 Hydroxysteroid (17-beta) dehydrogenase 6	
<pre>image: image: imag</pre>			
<pre>He_17437g Cyp2d10 Cyrochrome P450, family 2, subfamily d, polypeptide 10 He_26741 Uyt2D1 UD Fplouroschyltransferaez 2 family, polypeptide 51 He_340435 RepLP Riboscami protein, large, P2 He_37144 RepLP Riboscami protein, large, P2 He_37145 RepLP Riboscami protein, large, P3 He_37145 RepLP Riboscami protein, large, P4 He_37145 Riboscami protein, large, P4 He_37145 Riboscami protein large, P4 He_37145 Riboscami protein RepLP Riboscami protein He_441134 Ces2 Carboxylesterase 2E He_17057 Histifie Historic Cluster 1, Hie He_37145 Riboscami protein 1 He_321453 Robecami brotein 1 He_321533 Acall Acyl-CoA synthetase long-chain family member 1 He_331634 Gmil218 Riboscami protein 1 He_331634 Gmil218 Riboscami protein 1 He_331634 Gmil218 Riboscami protein 1 He_331635 Upperlise Directoryltransferaez 2 family polypeptide 55 He_331635 Upperlise Directoryltransferaez 2 family polypeptide 55 He_322805 P416 Frotein disulfid isomerase associated 6 No. 000000 DefL Directoryltransferaez 2 family of antiperlise 1 He_322805 P416 Frotein disulfid isomerase associated 6 No. 000000 DefL Directoryltransferaez 2 family of antiperlise 5 He_322805 P416 Frotein disulfid isomerase associated 6 No. 000000 DefL Directoryltransferaez 2 family of antiperlise 5 He_330184 Contactoryltransferaez 2 family of antiperlise 5 He_322805 P416 Frotein disulfid isomerase associated 6 No. 000000 DefL Directoryltransferaez 2 family of antiperlise 5 He_322805 P416 Frotein disulfid isomerase associated 6 No. 000000 DefL Directoryltransferaez 2 family of antiperlise 5 He_330184 Contactoryltransferaez 2 family of antiperlise 5 He_330184 Contactoryltransferaez 2 family of antiperlise 5 He_330184 Contactoryltransferaez 2 family of antiperlise 5 He_330184 Cont</pre>		Mm.20770 Cyp2al2 Cytochrome P450, family 2, subfamily a, polypeptide 12	
<pre>Ma.2011 Up;210 UDF glouuronceyltransferase 1 family, polypeptide B1 Ma.30120 MiceS MD repeat domain 89 MiceS MD repeat do</pre>		Mm.174372 Cyp2d10 Cytochrome P450, family 2, subfamily d, polypeptide 10	
<pre>Mail Wates W0 repeat domain 89 Bm_30433 Rpip/ Riboscama (protein, large F2 Bm_37134 Rpip/Riboscama (protein, large F2 Bm_37354 Rpip/Riboscama (protein, large F2 Bm_37354 Rpip/Riboscama (protein, large F2 Bm_37344 Rpiate Frocteni disulfide (informate associated 4 Bm_37354 Rpip/Riboscama (protein) Riboscama (protein) Bm_3744 Rpiate Frocteni disulfide (informate associated 4 Bm_37354 Rpip/Riboscama (protein) Riboscama (protein) Bm_3744 Rpiate (protein) Riboscama (protein) Bm_3744 Rpiate (protein) Riboscama (protein) Bm_3745 Riboscama (protein) Bm_3745 Riboscama (protein) Ribosca</pre>		Mm.26741 Ugt2b1 UDP glucuronosyltransferase 2 family, polypeptide B1	
<pre>Apple Alboard Apple Alboard protein, large F2 bits_2017 Apple Alboard protein Apple Alboard protein Abbar (5651507 5', mSNA sequence bits_2017 Opple Alboard protein Alboard protein Abbar (5012) (5', mSNA sequence bits_2017 Opple Alboard protein Alboard protein Abbar (5012) (5', mSNA sequence bits_2017 Opple Alboard protein Alboard protein Abbar (5012) (5', mSNA sequence bits_2017 Abbar (5') (5') (5') (5') (5') (5') (5') (5')</pre>		Mm.341719 Wdr89 WD repeat domain 89	
Rep. 2016 Sp. 2017 Hep. 501 Hear Shock Torkin, Large, DO Dim. 5777 Hep. 501 Hear Shock Torkin (School Hard Hear Control Hard Hear Control 122244402 ACENCOURT 101010514 NIM MCC 148 Num surculus CDNA clone INAGE:6561507 5', mENA sequence He. 2442 Feitaf Forcient disulfide incomerce associated 4 He. 373575 Cyp2054 Cytochrome P450, Family 2, subfamily c, polypeptide 54 gi122374450 ACENCOURT 801005 NIM MCC 128 Num surculus CDNA clone INAGE:6511216 5', mENA sequence He. 441136 Ces2e CarboxyLesterase 2E Ba. 170567 Histinthe Histone cluster 1, Hie He. 13705 Rep. Rhopi Alboxeme binding protein 1 He. 210352 Aceil Acyt. CoA synchetase long-chain family member 1 He. 210353 Aceil Acyt. CoA synchetase 10ng-chain family member 1 He. 23055 Uptochrome bis Feductase 3 He. 39158 UptochromyLingeride transfer protein He. 23055 Uptochrome bis Feductase 2 He. 39158 UptochromyLingeride transfer 2 family, polypeptide 85 He. 228055 Uptochrome Diffe SomeLing, polypeptide 85 He. 228055 Didd Frotein disulfide isomerase associated 6 Num 0000000 Diffe Distributer 2 family of the transfer 2 family of the transfer 2 family of the transfer 2 family and the some 2 family family for the transfer 2 fami		Mm.300433 Kplpz Ribosomal protein, large P2 Mm.201545 Dubosomal protein, large P2	
<pre>ima_0772 HepGobl Heat shock protein 90, beta (05p94), member 1 1129244682 ARCHORT 10195514 HW MCG 143 New musculus CORA clone THAGE:6561507 5', mRNA sequence 18.2442 Folder disulfide isomerase associated 4 18.37925 Opple46 Cytochnows 4950, fmail; y, subfamily c, polypeptide 54 31122371429 ARCHORT 1019514 Histore 119 Mus musculus CORA clone TMAGE:6511216 5', mRNA sequence 18.13705 Histore 1010 HTM MCC 119 Mus musculus CORA clone TMAGE:6511216 5', mRNA sequence 18.13705 Histore 1010 HTM MCC 1010 HTM MCC 1010 HTM MCC 1011216 5', mRNA sequence 18.13705 Histore 1010 HTM MCC 1010 HTM MCC 10112 HTM HEAT 1 18.13705 Histore 110 Housean 1 Histore 1 18.13715 Hubleolan 18.231134 Gail218 Ribosean Protein 113 peudogene 18.23129 Hubleolan 18.222205 Fild6 Frotein disulfide isomerase associated 6 19.000700 Filds Filds Frotein disulfide isomerase associated 6 19.000700 Filds Filds Frotein disulfide isomerase associated 6 19.000700 Filds F</pre>		Mm.5286 Robot Robot Protein, large, PO	
01122046402 ADENCOURT_10108514 NIM NGC_143 Num musculus CDNA clone INAGE:661807 5', mRNA sequence 082442 Fdiat Frocient intuitide internate associated 4 18373575 Cyp2054 Cytochrome P450, Family 2, subfamily c, polypeptide 54 01122371420 ADENCORT_601005 NIM NGC 128 Num succlus CDNA clone INAGE:6511216 5', mBNA sequence 18441136 Ces2e Carboxylesterase 2E 1810705 Ristlink Histone cluster 1, Hie 1810705 Ristlink Microscol Lineater 1, Hie 1820232 Acel 14014 1820233 Acel 14014 1820235 Discomal Toto 64014		Mm.87773 Hsp90bl Heat shock protein 90, beta (Grp94), member 1	
Ha. 2442 Folia disulfide information associated 4 Ha. 379375 Cyp2c63 Cytochocome 4950, family 2, subfamily c, polypeptide 54 g1122371425 AGENCOURT 801205 NIH MCC (129 Mus musculus cDNA clone HAGE:6311216 5', mRNA sequence Ha.41136 Cere Carboxylesterase 2E Ha.170587 Histihie Histone cluster 1, Hie Ha.154378 Noi Nucleolin Ha.154378 Noi Nucleolin Ha.210352 Acel Acyl. Cock synthetase long-chain family member 1 Ha.220523 Acel Acyl. Cock synthetase long-chain family member 1 Ha.231535 Ogtochocome bis Feductase 2 family, polypeptide 85 Ha.231535 Ogtochocome professional protein dispersion of the family for the family for the family for the family family for the family for the family for the family family for the family family for the family family family for the family family for the family fa		gi 22844682 AGENCOURT 10198514 NIH MGC 143 Mus musculus cDNA clone IMAGE:6561507 5', mRNA sequence	
<pre> the.379375 Cyp2045 Cytochrome P450, Family 2, subfamily c, polypeptide 54 gi12371425 ASKNOWT 801020 NHI MoC 129 Num seculus CDNA clone HAGE:6511216 5', mBNA sequence ha.441136 Ces2 Carboxylesterase 2E Ba.10705 Histinith Histone cluster 1 Hie Ba.20052 Acsil Acyt-CoA synchetase long-chain family member 1 Ha.20053 Acsil Acyt-CoA synchetase 1 Generate 2 family polypeptide B5 Ha.3075 Ugt208 Clubesman First Acyt-Complexater 2 family polypeptide B5 Ha.202825 Filia6 Frotein disulfide isomerase associated 6 Ym.0007000 During Direct Acytone 2 family a subfamily a filiant for the term of term of term of the term of the term of term</pre>		Mm.2442 Pdia4 Protein disulfide isomerase associated 4	
<pre>gi122371425 AGENCOURT_BOILOS NIH NGC 129 Mus musculus cDNA clone HGAGE(6311216 5*, mBNK sequence bm.41136 Ces2C Carboxylesterase ZE bm.170507 Ript Ruboxem binding protein 1 bm.19478 Not Nucleolin bm.210323 Acsil Acyl-CoA synthetae long-chain family member 1 bm.22050 CytoSch VycoArome bis Feductaes 2 family, polypeptide B5 bm.371854 Gmi2718 Ruboxema binde strokytower associated 6 bm.222823 Fala& Frotein dimilified isomerase associated 6 bm.20040 CytoSch Company Content and CytoSch Cyto</pre>		Mm.379575 Cyp2c54 Cytochrome P450, family 2, subfamily c, polypeptide 54	
18. 441145 Cep2 Carboy/sectrase 28 18. 10407 Mistlife Miscone cluster 1, Nie Miscone 2000etan 1 18. 10407 Mistlife Miscone 2000etan 1 18. 20032 Miscone 2000 Miscone 2000etan 1 18. 20032 Accil Acyl. Code synchecase long-chain family member 1 18. 20032 Accil Acyl. Code synchecase long-chain family member 1 18. 20032 Accil Acyl. Code synchecase long-chain family member 1 18. 20032 Accil Acyl. Code synchecase 2 family number 1 18. 20032 Miscone 2000 Miscone 2 family number 1 18. 20032 Folds Forcein disulfide isomerase associated 6 18. 20032 Folds Forcein disulfide isomerase associated 6 18. 20032 Folds Forcein disulfide isomerase associated 6		gi 22371425 AGENCOURT_8801205 NIH_MGC_129 Mus musculus cDNA clone IMAGE:6311216 5', mRNA sequence	
<pre>imiliar in the first one cluster 1, ife bm.17058 rtpl Riboare binding protein 1 bm.18478 Ncl Nucleolin bm.27032 Ncl Nucroscal riglyceride transfer protein bm.27032 Acoll Acyl-CoA synthetase long-chain family member 1 bm.270350 Cytobares bis Seductase 2 family, polypeptide B5 bm.272825 File& Protein dimiliar is seductase 2 family, polypeptide B5 bm.222825 File& Protein dimiliar is seductase 1 for the seductase 1 for th</pre>		Mm.441134 Cesze Carboxylesterase 2E	
18.19470 Nol Nucleolin 18.2941 Ntcp Nucreomal Triglyceride transfer protein 18.20233 Acsil Acyl-CoA synthetase long-chain family member 1 18.202506 Cybbs Seductase 3 18.391834 Gml2918 Ribosomal protein li3 preudogene 18.202757 Ugr285 UD0 glouronopyltransferae 2 family, polypeptide 85 18.222825 PdiaG Frotein disulfide isomerase associated 6 Nucleon December 2016 Semilural antipoly of content of a 19.202082 DdiaG Frotein disulfide isomerase associated 6		Mm.170587 Histinie Histone Cluster 1, Hie Mm.12705, Ethni Bibacene biding protein 1	
March March March Marchael (19) Verified transfer protein Marchael Loyi-Code synchrone b5 reductate 3 Marchael Loyi-Code Synchrone b5 reductate 3 Marchael Marchael Loyi-Code Synchrone b5 reductate 3 Marchael Barboard Torotein L13 preudogene Marchael Synchrone State Torotein State Synchrone B5 Marchael B Forcein State Synchrone B55 Marchael B Forcein State Synchron		Mm.15705 REDFI REDSOME DINAING PROCENT	
18.210232 Acsil Acy2-CoA synchetase long-chain family member 1 18.22606 CytoSto Synchotnem bis reductase 3 18.391834 Gml2918 Ribosomal protein Li3 preudogene 18.291975 Ugr2000 Diversional resolutione 2 family, polypeptide 85 18.222825 PdiaG Protein disulfide isomerase associated 6 19.0000000 Duran Diversional Action 2 acti		Mm.2941 Mttp Microsomal triglyceride transfer protein	
Han.22560 Cyb513 Cytochrome b5 reductase 3 Han.301834 Gmil218 Ribosmal protein L13 preudopene Han.291375 Ugr2D5 UDF glouuronosyltransferase 2 family, polypeptide 85 Han.222222 Filia Forcian injuridid insomrare associated 6 The 00/2000 Chinacharame Diff. Semila. A subdemily d alternative 0 Chinacharame Diff. Semila. A subdemily d alternative 0		Mm.210323 Acs11 Acy1-CoA synthetase long-chain family member 1	
Ha. 391834 Gal2518 Ribosomai procesn li3 preudogene Ha. 291575 Ugr2505 UGP glouronopyltransferaes 2 family, polypeptide BS Ha. 222825 PdiaG Procein disulfide isomerase associated 6 Control Control Contro		Mm.22560 Cyb5r3 Cytochrome b5 reductase 3	
Ha. 291575 Ugr2DS UDF glucusconsyltransferase 2 family, polypeptide 85 Ha. 222257 Folick Forcian isolatified incomerce associated 6 (M. 000700 Output Characterian State Sample 2 and Sample 2 and Sample 2 (M. 000700 Output Characterian State Sample 2 and Sample 2 (M. 000700 Output Characterian State S		Mm.391834 Gm12918 Ribosomal protein L13 pseudogene	
im.22202 Full& Frotein daulide isomerse associated 6		Mm.291575 Ugt2b5 UDF glucuronosyltransferase 2 family, polypeptide B5	
e		Mm.222022 False Frotein disulfide isomerase associated 6	
	< [

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.

2. Long Language (JPL2004) ER Services Charles Forder	
Tex [12] A set (Marco) Definition (Constraints)	
+ S N Internet Sector S	PRIMI and generated METVIC control AND 🗇 🗸 🖓 - Graph P
👔 Maat Valand 😻 Gering Statest 🐞 Latest Handlines 🚼 Google 😇 Overview Use 2 Part	Warld Clock - TL. 😨 Overview (Java Platfor 🗌 Matrix Science - Home 🔝 Matrix Science - Datab
p1114411002 Means 20042 Source 220 Matchess 18(18) Sept 60290294071 NCT_COLP_Mean Test Actions 200407(502100) S 40290294071 NCT_COLP_Mean Test Actions 200407(502100) S	nnoes: 4(3) Prame: 2 , mBCG sequence
10. 1111111111111111111111111111111	termine trib match : Life France : L control to match : Match termine The Second Sec
Setet Summay Report (PR02008 657+	1
🔶 🕭 📡 🗋 begeng inascot, 2,4,0,64/ cg/imaster, results pl?files - 527 data %3720, 2042/ LARX, gueryfists all&REPTVPE-select&URX 🏠 🗢 C 🛃 - Google 🖉 🚱	ee: 7(4) is cDHL 5', wHHL sequence
🙆 Most Visited 🦉 Getting Started 🏊 Latest Headlines 👌 Google 🖸 Overview (Java 2 Platf., 🗠 The World Clock – TL. 🧟 Overview (Java Platfor., 🗌 Matrix Science - Home 🗌 Matrix Science - Datab	141 7(4) Frame: 2
23907 866.6286 1735.2427 1734.8402 0.4025 0 61 0.03 1 U R.VOREAQCLVRELE.R.	m: 7(4) Frame: 2
26/36 945.1098 1932.5446 1932.0772 0.4674 0 60 0.032 1 0 R.GTTVITSLSSVLRDSR.E 26747 26756 26751	is cDDA 5', sR3DA sequence Ma: 7(4)
	is cOlk 5', mAUA sequence
7. <u>Un:16660</u> Soore: 398 Matches: 22(11) Sequences: 11(5)	ne: 5(4) em2A2: 1.02 Frame: 1
wind Froigh F-sydroxylass, sets polymption Ouerv Observed Mr(excl) Mr(calc) Delta Miss Score Except Rank Unione Peptide	B misculus cDNA close NEA/A0312A06 ENAGE:20734309 5", mBDBA sequence Enk Unique Peptide
7649 575.9399 1149.8653 1149.6637 0.2016 0 52 0.42 1 U K.ALAPEYAK.A	1 R.SHFEGEITK.E 19920
7952 579.9505 1157.8865 1157.5808 0.3057 0 71 0.0046 1 0 K.SVSDYDOK.L 7946	B. LITLEEDITE. V 18546 18549 18555
11010 628.0115 1254.0085 1253.7627 0.2458 0 53 0.28 1 U R.ILEFFOLK.K	L
11133 629-5551 1257-6957 1257-7325 0.1632 0 54 0.22 1 0 K.QLAPINOK.L 1122	
14177 706.001 1413.969 1414.765 -0.776 0 51 0.42 1 0 R.NETERLINK Y	
18442 745.8950 1489.7754 1489.8020 -0.0266 0 67 0.01 1 U B.EADDIVENIE 18452 18456	
18550 748.0871 1494.1596 1493.8254 0.3342 0 83 0.00025 1 U R.IITLERENTK.Y 18546 18549 18549 18555	NATE data, guardata all'ATENTA colonalistati 🕆 = C 🚺 - Google 🖉 🖉
18790 502.3231 1503.9474 1503.7925 0.1549 0 55 0.13 1 U K.SNFEEALAANK.Y	and Check - TL. 😨 Overview Dava Platfor. 🗌 Matrix Science - Home 🗌 Matrix Science - Databa
18970 756.0795 1510.1444 1509.8204 0.3241 0 (56) 0.11 1 U R.IITLEERSTR.Y 18963	1' state annual a
25372 614,7594 1841,2563 1840,9783 0.2779 1 60 0.03 1 0 K.YEFESDELTAEK.I	
	1 5', sRUG sequence
9. 20.6696 Score: 388 Natches: 22(13) Sequences: 7(6)	Daigne Peptide 0 K.LTELEVAR.V
Rdh7 Retinol dehydrogenase 7	E. FPLTTERAME. E 14118
Query Observed Mr(expt) Mr(calc) Delta Miss Score Expect Rank Unique Peptide	3(1) emPAT: 0.57
9/92 393.3202 1104.6239 1104.0059 0.0139 0 90 0.033 1 0 R.YSAADAK, 15794 620 5521	s302 sequence
12120 644.4271 1286.8396 1286.7114 0.1282 0 54 0.19 1 U K.MFLDSYKA, 12124	R.ALAFEYAR.A
15346 694.4422 1386.8698 1386.6901 0.1797 0 67 0.0076 1 U K.YOYZAFSDSLR.R	0 B.ATTERALAARY.Y
16296 710.0087 1418.0029 1417.8636 0.1393 0 65 0.016 1 U R.LETVILOVTK.T	A COMPANY OF AN
18205 741.0224 1480.0302 1479.7813 0.2489 0 68 0.0067 1 U R.1M00TSSEVR.8 18200 18201 21664 811.2202 1620.4259 1619.9126 0.5133 0 75 0.0013 1 U R.TESIVAATQWW.8 21674 21675 21675 21675 21675 21675 21674 21674 21675 21675 21675 21674 21675	(3(1) emPAI) 0.63 Frame: 3
	Unique Peptide
	R. VQ0EIDAVIDQVR.C U K.GTILIPHOKSHNIK,D 20112 20114 20114 20119
24157 555,5429 1754,5428 1755,5921 0.5748 1 56 0.1	1 B. HV0081DAV100VR.C
122 <u>01154439</u> Massi 21994 Source 135 Matchess 3(2) Sequence matching. In Life Technology and the International Source Conference on Confer	es: 5(2) emBAI: 0.55 Frame: 3 DNA close DMADE:415083 5' similar to ub:204877 SERUM ALMONIM PRECURSOR (MIMAN):, mJ
Quary Generated Mringell Hrinkland Bulla Has Source Expect	Rank Unique Peptide
2017 793-074 1553-140-157 1527-1630 0.4655 0 0 1 0.4	1 0 E.ARGVETPTLVEAAR.H
2 <u>9442</u> 753.5987 2257.7743 2254.0915 -0.3172 1 60 0.025	1 0 R. AADRDYCPSTEOPHLVYR. C
Mg///hoppray/massed, 2, 4, 5, 54/ w/wmiane view.adfl	iena?
and a second secon	
MASCOT : Nucleic Acid Databases	Science (MATRIX)
	(SCIENCE)

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.

	Mus musculus (10 52) - Genom File Edit View Higtory Bookr	e - NCBI - Mozilla Firefox narks <u>I</u> cols <u>H</u> elp						
	S Mus musculus (ID 52) - Genom	e - NCBI +						
	• • • • • • • • •	w.ncbi.nlm. nih.gov /genom	e/iterm=mus musculus		-		☆ ▼ C Google	P 💽 🚖
	Most Visited ¹ / ₂ Getting Starte	ed 📄 Latest Headlines 🚼	Google 🖸 Overview (Java 2 Platf 🎌	The World Cl	iock – Ti 🖸 Overview (Java Platfor 🗌 Matrix S	Science - Home 🗌 Matrix Scie	ence - Datab	
		S NODI RESOURCE					My NCBI Sign In	1
		Genome	Genome 👻 mus mus	culus			Search	
			Save searc	h Limits	Advanced		Help	
		Display Settings: 🕑	Overview			Send to: 🕑		
							NCBI Resources	
			lus musculus (house e laboratory mouse is a major model or	e mous Iganism for b	e) asic mammalian biology, human disease, and ge	enome evolution, and its	Genome Reference Consortium (Mouse) Map Viewer	
		94 94	nome has been sequenced					
		Lir Eu	eage: Eukaryota[1451]; Metazoa[587]; Chi archontoglires[57]; Glires[26]; Rodentia[2	ordata[245]; Ci 4]; Sciurognat	aniata[240]; Vertebrata[239]; Euteleostomi[234]; Ma hi[20]; Muroidea[14]; Muridae[4]; Murinae[4]; Mus[2	immalia[113]; Eutheria[109]; []: Mus[2]; Mus musculus[1]	Tools	
		The mouse is one of t providing a wealth of o	he major organisms for modeling huma different genotypes and phenotypes for	n disease an genetic and o	d comparative genome analysis. There are over other studies. In addition, thousands of spontane	450 inbred strains of mice, eous, radiation- or	BLAST Genome	
		chemically-induced, a	ind transgenic mutants provide potentia	I models Mo	re		Publications	
		Organism Overv	iew See also: Genome list Organe	lle List			Intragenic enhancers act as alternative	
			Chromosomes		Assembly and Annota	ation	promoters. [Mol Cell 2012]	
		1 2 3 4 5	6 7 8 9 10 11 12 13 14 15 1	6 17 18 19	Default assembly	v	stem cell function through dere [Genes Dev 2012]	
					5 other assemblies are a	vailable	Deep-sequencing identification of the genomic	
		BBeder	3086666 <u>6</u> 0 8 0.		Assembly Name	MGSCV37	targets of the cytidine deamin [Nat Immunol 2011]	
			ě ě		Highest level of assembly	Chromosome	See more	
					Size (total bases) Number of genes	2,716,949,182 34,256		
			Related BioProjects		Number of proteins	27,051	BioProject	
			Type	Caunt	Mitochondrial Geno	me	Gene	
		RefSeq Genome			Last record update	21-0ct-2010	Protein Clusters	
		Clone ends			Last sequence update Size	16,299	Components	
		Epigenomics		34	Number of genes	37	Protein	
		Genome sequencing		1	Promoter or proteins	13	PubMed	
		Map					Taxonomy	
		Other		12				
		Targeted Locus (Loci)					Search details	
		Transcriptome or Gene	expression	600			"Mus musculus"[Organism]	
		Variation		6				
							Search See more	
							External Resources	
							Broad Institute	
							CIDR Mouse Genotyping	
							Celera	
							EU RH map	
							Ensemhl Genome Browser	-
M	ASCOT	: N	ucleic Acid	l Da	tabases	© 2012	Matrix Science	{MATRIX} {SCIENCE}

You can download the mouse genome sequences from the NCBI.

Mem Serve: Detal.	A mutsculus/Assembled_chror A_musculus/Assembled_chror Size 5005 8 5005 8	In the second secon
Last Modified 040/07/02 22:800 040/07/02 22:800 040/07/07/07/07/07/07/07/07/07/07/07/07/07	4 musculus/Assembled_chror Size 510.05 310.05 300.0	Index of ftp://tp.ncbi.nlm.nih.gov/genomes/ to to higher twi directory Mark Mark Mark Mark Mark Mark Mark Mark Mark
East Modified 001/002 238/8 001/002 238/8 001/00	A_musculus/Assembled_chron Size אור איז איז איז אור איז	Index of ftp://tp.ncbi.nlm.nih.gov/genomes/ to higher level develop: Mane Mane Mane Mane, http://doc.doc.got.doc.got man.gt/bm/clane.gt/bm/clane.gt man.gt/bm/c
Amosomes/seq/ Last Matified (#00/2012 1238/8 (#00/2012 1238/8	A_musculus/Assembled_chror Size 3025 8 3080 9 3080 9 300 9 3000 9 30000000000	Index of ftp://tp.ncbi.nlm.nih.gov/genomes/ to the higher lead at setup: Harm Harm Harm, Lifthon, Colon, colif.org Harm, Lifthon, Colon, colif.org, Stat. Harm, Lifthon, Colif, Colif, Stat. Harm, L
Last Modified 00/07/02 23:00 00/07/02 23:00	■ mosculus/esseniores_chrones Size	Index on Index and Index of In
Last Modified 06/07/07 22 23/06 06/07/07 23/07 06/07/07 23/06 06/07/07 23/06 06/07/07 06/07/07 06/07/07 06/07/07 06/07/07 06/07	Size 5423 08 98444 08 9809 48 9809 48 9809 48 9809 48 9809 48 9804 48 9804 48 9804 48 9804 48 9805 48 9805 48	To be the final field entering: Name With an angle Allow Colours, and A angle With angle Allow Colours, and Allow Gold Allow Colours, and Allow Gold Allow Gold Allow Colours, and Allow Gold Allow Colours, and Allow Gold Allow Gold Allow Colours, and Allow Gold
Last Modifield Weinyma 2an We	Size 5425 50 8044 60 3031 60 3031 60 3032 60 3032 60 3032 60 3032 60 3044 60 3044 60 3045 60 3045 80 3045 80 3045 80	Name Simple State State Name Simple State State State Simple State State State State Simple State State State Simple State State State Simple State State State State Simple State State State State Simple State State State State State Simple State State State State State Simple State State State State State State Simple State S
96/03/22 22:80/6 96/03/22 22:80/6	54125 30 3014 64 3019 30 3019 30 3049 40 3040 40 30 3040 40 30 30 30 30 30 30 30 30 30 30 30 30 30	man, k, k, hard, configured and app man, k, k, hard, k,
949(1)/22 (2)/28/9 949(1)/22 (2)/28/9 949(1)	541.45 KB 3601 KB 31027 KB 31027 KB 30603 KB 30603 KB 30603 KB 32554 KB 34554 KB 34554 KB 34557 KB 34557 KB 34557 KB	mon_kit/m_colour_colour_colour_performance mon_kit/m_colour_colour_colour_performance mon_kit/m_colour_colour_colour_performance mon_kit/m_colour_colour_colour_performance mon_kit/m_colour_colour_colour_performance mon_kit/m_colour_colour_colour_performance mon_kit/m_colour_colour_colour_performance mon_kit/m_colour_colour_performance mon_kit/m_colour_performance mon_kit/m_
840/07/2012 22:30/86 940/07/2012 22:30/86	96031 48 38939 48 31317 48 3544 48 30031 48 30021 48 3021 48 34424 48 3271 48 3271 48 34429 48	 mmulti Munch Carlow, Michael Anger
99010/302 22.20980 99010/302 22.20980 99010/302 22.20980 99010/302 22.20980 99010/302 22.20980 99010/302 22.20980 99010/302 22.20980 99010/302 22.20980 99010/302 22.20980	300/97 KB 33121 KB 35444 KB 30063 KB 30021 KB 34644 KB 34644 KB 34644 KB 34645 KB 34645 KB	 mm, 24, 26, 45, 46, 45, 46, 46, 46, 46, 46, 46, 46, 46, 46, 46
00101202 22300 0010202 22300 0010202 22300 0010202 22300 0010202 22300 0010202 22300 0010202 22300 0010202 22300 0010202 22300	3544 KB 30653 KB 30202 KB 34834 KB 34834 KB 34639 KB 34629 KB 34629 KB	 mon, all, thou, Colona, child Lindiaga mon, all, thou, Colona, child Lindiaga mon, all, thou, Colona, child Lindiaga mon, all, thou, Colona, child Lindiga
001/03/2 2280 001/03/2 2280 001/03/2 2280 001/03/2 2280 001/03/2 2280 001/03/2 2280 001/03/2 2280	30863 KB 3022 K0 32543 KB 34254 KB 32371 KB 34629 KB 38627 KB	 mm_ikit/im_Colors_chtt2Lagt mm_ikit/im_Colors_cht2Last mm_ikit/im_Colors_cht2Last mm_ikit/im_Colors_cht2Last mm_ikit/im_Colors_cht2Last mm_ikit/im_Colors_cht2Last mm_ikit/im_Colors_cht2Last
99/09/2012 22:30/00 99/09/2012 22:39/00 99/09/2012 22:39/00 99/09/2012 22:39/00 99/09/2012 22:39/00 99/09/2012 22:39/00 99/09/2012 22:39/00 99/09/2012 22:39/00	30021 KB 32543 KB 34634 KB 32371 KB 34629 KB 26878 KB	 mm_abl_Mm_Colors_0h02.ordingt mm_abl_Mm_Colors_0h03.htgp mm_abl_Mm_Colors_0h03.htgp mm_abl_Mm_Colors_0h04.htgp mm_abl_Mm_Colors_0h04.htgp mm_abl_Mm_Colors_0h04.htgp
0001/0021 222800 0001/0021 222800 0001/0021 222800 0001/0021 222800 0001/0021 222800 0001/0021 222800 0001/0012 222800	32543 KB 34834 KB 32371 KB 34629 KB 26878 KB	■ mm_alt_Mm_Colors_chd3ds.gt ■ mm_alt_Mm_Colors_chd3ds.gt ■ mm_alt_Mm_Colors_chd4ds.gt ■ mm_alt_Mm_Colors_chd4ds.gt ■ mm_alt_Mm_Colors.dtd.and.gt
00/03/2012 222800 00/03/2012 222800 00/03/2012 222800 00/03/2012 222800 00/03/2012 222800 00/03/2012 222800	34834 KB 32371 KB 34629 KB 28878 KB	 mm_alt_Mm_Celera_chrl3.mla.gz mm_alt_Mm_Celera_chrl4.fa.gz mm_alt_Mm_Celera_chrl4.mla.gz
09/03/2012 22:29/0 09/03/2012 22:29/0 09/03/2012 22:29/0 09/03/2012 22:29/0 09/03/2012 22:29/0 09/03/2012 22:29/0	32371 KB 34629 KB 28878 KB	mm_sit_Mm_Celera_chrl4.fa.gz mm_sit_Mm_Celera_chrl4.mla.gz
09(03)2012 22:29:00 09(03)2012 22:29:00 09(03)2012 22:29:00 09(03)2012 22:29:00	540,29 KB 28878 KB	min at this celeta chila da
09/03/2012 22:29:00 09/03/2012 22:29:00 09/03/2012 22:29:00	20070 MB	Compared the Colors shall be as
09/03/2012 22:29:00	20896 KB	mm at Mm Calera christelia
	27164 KB	mm_alt_Mm_Celera_chr16.fa.gz
99/93/2012 22:29:00	29073 KB	mm_alt_Mm_Celera_chr16.mla.gz
09/03/2012 22:29:00	24920 KB	mm_alt_Mm_Celera_chrl7/fa.gz
09/03/2012 22:29:00	26627 KB	mm_alt_Mm_Celera_chv17.mfa.gz
09/03/2012 22:29:00	25064 KB	mm_alt_Mm_Celera_chr18.fa.gt
09/03/2012 22:29:00	26831 KB	mm_alt_Mm_Celera_chr18.mfa.gz
09/03/2012 22:29:00	16682 KB	mm_att_Mm_Celera_chr29Ma.gz
09/03/2012 22/28/00	50402 KR	mm alt Mm Celera chr2 fa.gz
09/03/2012 22:28:00	53910 KB	mm_alt_Mm_Celera_chr2.mfa.gz
09/03/2012 22:29:00	45109 KB	imm_alt_Mm_Celera_chr3.fa.gz
09/03/2012 22:29:00	48250 KB	mm_alt_Mm_Celera_chr3.mfa.gz
09/03/2012 22:29:00	41876 KB	mm_alt_Mm_Celera_che4.fa.gz
09/03/2012 22:29:00	44800 KB	mm_alt_Mm_Celera_che4.mfa.gz
09/03/2012 22:29/00	40350 KB	mm_alt_Mm_Celera_chr5.fa.gz
09/03/2012 22:29/00	43193 KB	mm alt Mm Celera chef face
09/03/2012 22:29:00	44372 KB	mm_ah_Mm_Celera_chr6.mfa.gz
09/03/2012 22:29:00	36993 KB	mm_alt_Mm_Celera_chr7.fa.gz
09/03/2012 22:29:00	39575 KB	mm_alt_Mm_Celera_chr7.mda.gz
09/03/2012 22:29:00	35570 KB	mm_alt_Mm_Celera_chr8.fa.gz
09/03/2012 22:29:00	38053 KB	mm_alt_Mm_Celera_chr8.refa.gz
09/03/2012 22:29:00	33866 KB	mm_alt_Mm_Celera_chr9.fa.gz
09/03/2012 22:29:00	36270 KB	e mm_at_htm_celera_ches.mta.gz
09/03/2012 22:29:00	41776 KB	mm.alt. Mm. Celera. chrX.mla.gz
09/03/2012 22:29:00	538 KB	mm_alt_Mm_Celera_chrV.fa.gz
09/03/2012 22:29:00	574 KB	mm_alt_Mm_Celera_chrV.mla.gz
		··· · · · · · · · · ·

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

Select Summary Report (IPR020	8 Mous. + na/mascret 2.4.0.61/cai/master results alffiles. %2Fdata%2F0012020F001439.dat& auendict-all@8FDTVPF-celect& circl 🔿 🛛 🖉 - Google 🖉 🤷
A Most Visited 🥹 Getting Started	💫 Latest Headlines 🔧 Google 🧧 Overview (Java 2 Platf 💝 The World Clock – Ti 🧧 Overview (Java Platfor 🗌 Matrix Science - Home 🗌 Matrix Science - Datab
(MATRIX) SCIENCE) Masco	t Search Results
User Email	PAB
Search title	iPRG2008 Mouse_genome
MS data file	D:\iPRG2008\mgf\merged.mgf
Database	Mouse_genome_genome_goll20123 (1362828 sequences; 5505027542 residues)
Enzyme	30 Apr 2012 at 13:56:22 GMT
Fixed modifications	ITRAO4Diex (K), ITRAO4Diex (N-term), Methylthio (C)
Variable modifications	Acetyl (Protein N-term), Gln->pyro-Glu (N-term Q), Oxidation (M)
Mass values	Monoisotopic
Protein Mass	Unrestricted
Peptide Mass Tolerance	± 0.9 Da
Fragment Mass Tolerance:	± 0.6 Da
Max Missed Cleavages	
Number of meries	53101
Protein hits	<u>gi 1372099100 ref [KC 00006.71 209</u> base 3474600-34776120 Mus musculus strain C578L/45 formconceme 6, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 1573</u> base 18544001-38156120 Mus musculus strain C578L/45 chromosome 6, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 0775</u> base 1972003-75924121 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099101 ref [KC 000074.61 0775</u> base 10248001-16576120 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099101 ref [KC 000074.61 0775</u> base 10248001-15786122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099101 ref [KC 000074.61 0775</u> base 10248001-15786122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099101 ref [KC 000074.61 0775</u> base 10248001-15786122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 0775</u> base 10248001-15786122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 0775</u> base 10248001-15786122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 0775</u> base 10258001-10256121 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 0775</u> base 10592001-10460122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 8755</u> base 105492001-10460122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 8755</u> base 105492001-104748122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 8755</u> base 105492001-104748122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 8755</u> base 105492001-104748122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 8755</u> base 105492001-104748122 Mus musculus strain C578L/45 chromosome 5, GRCM8 C578L/45 <u>gi 1372099100 ref [KC 000074.61 7745</u> base 10540001-2056121 Mus musculus strain C578L/45 chromosome 5,

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?

1.											
	g11372	2099108[ref	INC 000068.	71 2898	Mass: 173	010	score	629	Mate	nes: 2	(17) Sequences: 11(8) emPAI: 0.08
	Dases	34764001-3	Mr (amt)	Maculus s	Delta	DL/ CU	Saore	Expect	Bank	Jo Con	Pantida
	2720	488.3756	974.7367	974.6004	0.1363	0	54	0.36	1	onique	B. LTGDAAK, N
	9459	603,8705	1205,7264	1205.6747	0.0517	0	61	0.035		U	F VI PDCDI F F 0465
	11545	635,4900	1268,9654	1268.6856	0.2799	0	55	0.33	1	U	K. ETARAYLOK, K
	21354	803,1397	1604.2648	1603.8337	0.4311	0	63	0.023	1	U	D NELESVAVELE N 21353
	23465	855.9392	1709.8639	1709.8746	-0.0108	0	66	0.011	1	U	B. ITPSYVAFTPEGER 1. 23478 23479
	23473	571.0578	1710.1517	1709.8746	0.2770	0	(57)	0.088	1	U	R. ITPSYVAPTPEGER. L
	23594	859.6885	1717.3625	1717.8879	-0.5253	0	77	0.00079	1	U	R. TWNDPSVOODIK, F 23604
	25347	609.0556	1824.1449	1823.9946	0.1503	0	56	0.088	1	U	K. TFAPEBISAMVLTK.M
	27170	983.6427	1965.2709	1965.0047	0.2662	0	96	7.9e-06	1	U	K.NOLTSNPENTVFDAK, R 27165 27174 27175 27176
	28331	693.4307	2077.2703	2077.1078	0.1625	0	82	0.00019	1	U	K. DNHLLGTFDLTGIPPAPR. G 28351
	28790	1063.1951	2124.3757	2124.1306	0.2450	0	80	0.00027	1	U	K. SOIFSTASDNOPTVTIK. V 28786 28789 28793
	29891	755.2793	2262.8160	2262.1048	0.7112	0	69	0.0033	1	U	K. IEWLESHODADIEDEK. A 29889
	Query 19996 20427 20435 20855 20855	Observed 776.4102 784.4878 784.5320 792.4776 563.6700	Mr (expt) 1550.8058 1566.9611 1567.0493 1582.9406 1687.9882	Mr(calc) 1550.7166 1566.7115 1566.7115 1582.7064 1687.7942	Delta M 0.0893 0.2497 0.3379 0.2342 0.1940	155 S 0 0 0 0 0	core (75) 0 (70) (63) 77 0 (49)	Expect Ra .00097 0.0044 0.024 .00064 0.45	ank Ur 1 1 1 1	U U U U U U U	Feptide
3.	23054 g11372	845.1180	1688.2215	1687.7942 61 7073	0.4273 Mass: 462	0	92 2 Score	.8e-05	1 Mate	U hes: 10	VFANPEDCAGFOR.G 23055 (10) Sequences: 1(1) emPAI: 0.02 Frame: 2
	bases	84864001-8	4876120 Mus	musculus s	train C57	BL/6J	chron	tosome 18	, GRC	m38 C5	BL/6J
	Query	Observed	Mr (expt)	Mr (calc)	Delta I	Miss	Score	Expect 1	Rank U	Jnique	Peptide
	0041-		2149 1647	2349.0227	11.3420		(100)	2.1e-06	1	U	R. EOAGGDATENFEDVGHSTDAR, E 30427
	30413	11/3.009/	2040.0047								

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

Eile Edit View History	<u>Bookmarks</u> <u>Iools</u> <u>H</u> elp	
Peptide Summary Rep	irt (/data/201 × 🗍 GenBank format feature table × +	
< > S * [🛛 bogong/mascot_2_4_0_64/cgi/protein_view.pl?file=%2Fdata%2F20120501%2FF001462.dat&khit=gi]37209910 ☆ 🛡 🦿 Google	P 🚳 🟫
😬 Most Visited 👹 Gettir	g Started 🚵 Latest Headlines 🐒 Google 🧧 Overview (Java 2 Platf ∻ The World Clock – Ti 🧧 Overview (Java Platfor 📋 Matrix Science -	Home
BLASTCDS	complement(1120011223)	
	/label=Q97	
	/colour=2	
	<pre>/note="Mascot match, guery=97, mass=1086.48, score=31, rank=4, seguence=NYYEQWGK"</pre>	
	/blastp file="/data/20120501/F001462.dat"	
	/mass=1086.48	
	/score=31	
	/rank=4	
	/translation="NYYEQWGK"	
BLASTCDS	complement (1120011223)	
	/label=Q98	
	/colour=2	
	/note="Mascot match, query=98, mass=1086.48, score=31, rank=4, sequence=NYYEQWGK"	
	/blastp_file="/data/20120501/F001462.dat"	
	/mass=1086.48	
	/score=31	
	/rank=4	
	/translation="NYYEQWGK"	
BLASTCDS	complement(1120011223)	
	/label=Q99	
	/colour=2	
	/note="Mascot match, query=99, mass=1086.48, score=31, rank=4, sequence=NYYEQWGK"	
	/blastp_file="/data/20120501/F001462.dat"	
	/mass=1086.48	
	/score=31	
	/rank=4	
	/translation="NYYEQWGK"	
BLASTCDS	complement(1056510594)	
	/label=Q127	
	/colour=2	
	/note="Mascot match, query=127, mass=1187.64, score=69, rank=1, sequence=IDTIEIITDR"	
	/blastp_file="/data/20120501/F001462.dat"	
	/mass=1187.64	
	/score=69	
	/rank=1	
	/translation="IDTIFITIDR"	
BLASTCDS	complement (1056510594)	
	/label=gl28	
	/note="mascot match, query=128, mass=1187.69, score=69, rank=1, sequence=IDTIEIITDR"	
		CLC A CEDI

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.

Artemis Entry Edit: Chr6_51456001_51468120.fasta	
file Entries Select View Goto Edit Create Run Graph Display	
Selected Teature: bases 30 UI29 (/label=UI29 /colour=2 /mass=ll87.b4 /score=b9 /Fank=1 /translation="UTILLIDK" /note="Mascot match, query-	129, na
	I I 🖻
	CDS
CDS CDS	
10200 10300 10400 10500 10500 10700 10800 10900 11000 11100	11200 1
CDS 0222 CDS 0306 0129 CDS	
CDS CDS	CDS Q9: Q350 -
۹ ۲ ۲	
T K H T S * T D T Y R Q F Y P Q D H G H Q K # Q S Q S L F S F Q T A Y R # L F Q + Y Q F F H T I	Q S N L # G G VA
Y KAHLMNRHLQTILSTGSVSSKVTKPKPLFFPDCLSVIISIVSIFPYS	SK SLRWC
a caabee a contrate a cabe contrate a cabe a trittate cabe cabe contrate a a a contrate a cabe contrate cabe contrate a cabe contrate a cabe contrate cabe cabe contrate cabe c	TCAAAGTAATCTCTAAGGTGGTGT
LO480 LO500 LO500 LO520 LO520 LO540 LO540 LO550 LO550 LO540 LO550	10620
LACRMFLCRCVIGKDVFDHDDFTVF6F6	EFYDRLHH
V F C V E H V S V # L C N # G C S * P * * F Y C L U L R K E K U V A + R Y N N * Y Y * N K U V R	* L L R + P P T
CLVG*SCVSVSLKIWLINTNLLLLALAEKRELSGIPL#KLLILKENSP	LTIELTTN
BLASTOS 1055 10594 C HERO ELCH, QUAY127, EXECUTE, SCHEEN, KONCES, KUNCT, SCHEMCK CONCENTRATION BLASTOS 11200 11223 C Masco Hatch, QUEY97, Hassilo36.48, societa3, Kank-4, sequence-MYTEQNGK	Ê
BLASTEDS 11200 11223 c Massot match, query=98, mass=1086.48, soure=31, rank=4, sequence=NVTEQUGK	
BLASICDS 11224 1127 c Mascot match, query=39, mass=1797.91, score=70, rank=1, sequence=LFIGGLSFETTEESLR	
BLASTCDS 11224 11271 c Mascot match, query=350, mass=1797,91, score=70, rank=1, sequence=LFIGGLSFETTEESSR	
BLASIUDS 11224 112/1 c Mascot match, query-331, mass-1926.01, score=45, ruk=1, sequence=LFIGUSSFIIESSLR BLASICDS 11224 11274 c Mascot match, query-330, mass-1926.01, score=45, ruk=1, sequence=KLFIGUSSFITESSLR	
BLASTCDS 11224 11274 c Mascot match, query=371, mass=1926.01, score=45, rank=1, sequence=KLFIGGLSFETTEESLR	
ELASTCD3 11224 11274 c Mascot match, query=372, mass=1926.01, score=45, rank=1, sequence=KLFIGGLSFETTEESLR	
	
MASCOT : Nucleic Acid Databases © 2012 Matrix Science	SCIENCES

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.

P P F N II N	lass values rotein mass eptide mass tolerance ragment mass tolerance lax missed cleavages nstrument type umber of queries	: Unrestricted : ± 0.9 Da : ± 0.6 Da : 1 : ESI-TRAP : 33,191			
	Database	Size Avg	. 5% identity threshold	#matches > threshold	
			37	1910	
	UniProt_mouse_20120418	2.7 x 10^7	5,		
	UniProt_mouse_20120418 EST_mouse_20120429	2.7 x 10^7 4.5 x 10^9	60	465	
	UniProt_mouse_20120418 EST_mouse_20120429 Mouse_genome 20120123	2.7 x 10^7 4.5 x 10^9 2.7 x 10^9	60 60	465 367	

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

Incel Parket Record Protect Centre Processer	Entre	ez records	
160 as complete name • ? lock Ge Clear ay 3 levels using filter none •	Database name	Subtree links	Direct links
IS	Nucleotide	3,440	-
omy ID: 2706 ted blast name: eulicots	Nuclear de DOT	567.000	
genus (c code: Translation table 1 (Standard)	INUCLEOUDE EST	307,880	
hondrial genetic code: Translation table 1 (Standard)	Nucleotide GSS	46,391	-
les: Microcitrus	Destain	2 260	1
les: Fortunella	Protein	2,209	<u> </u>
ex Elementarius ity: Eremocitrus Swingle	Genome	1	
ity: Microcitrus Swingle	-		
ity: Fortunella Swingle	Popset	64	<u>61</u>
pr(full) cellular comismic Eukarvota: Veidelantae: Streptophota: Streptophota: Embrophota: EuphoRophota: Spermatophota: Magoskophota: readicorteledons: core eudicorteledons: roadi mobile: Straindole: Patacoae	GEO Datasets	190	3
anarona anarona	UniGene	24,930	-
	UniSTS	139	-
nents and References:	PubMed Central	1,024	331
lora of China - Rutaceae Xaurxianz Zhane, Thomas G. Hartlev & David J. Mabberlev, Rutaceae. In Wu, Z. Y., P. H. Raven & D. Y. Honz, eds. 2008. Flora of China, Vol. 11 (Oxalidaceae through Aceraceae). Science Pro	ss. Boiine. Gene	140	-
	SRA Experiments	8	-
bits to State Access and Access and Access and Access and Access and Access	Probe	195	-
ROPICOS (Jun 08, 2009) Name verified on 8 June 2009 in: Tropicos org. Missouri Botanical Garden, 2008 - 4344 Shaw Boulevard, Saint Louis, Missouri 63110, U.S.A. on line at. http://www.tropicos.org/	Bio Project	24	-
<u>VMI (Jun 08, 2009)</u> Same verified on 8 June 2009, in: The International Plant Names Index (2004). Published on the Internet http://www.ipri.org	Bio Sample	248	-
	Protein Clusters	70	
nal Information Resources (NCBI LinkOut)	-		
ut Subject LinkOut Provider	Taxonomy	117	<u>1</u>

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.



We have a nice example of a possible base insertion.