Alistair MacDougall[1], UniProt Consortium[1-4], Ensembl[1]

1 EMBL-EBI, Cambridge, UK; 2 Swiss Institute of Bioinformatics, Centre Medical Universitaire, Geneva, Switzerland
3 Protein Information Resource, Georgetown University Medical Center, Washington, USA; 4 Protein Information Resource, University of Delaware, Newark, USA

Mapping Genes to Proteins in UniProtKB

EMBL-EBI

UniProt Knowledgebase

The UniProt Knowledgebase (UniProtKB) endeavours to provide the scientific community with a comprehensive catalogue of protein sequence and functional information.

To make effective use of these data for genome studies, it is essential to have accurate mapping from gene to protein sequence, and in the reverse direction from protein to gene.

Mapping UniProtKB:Ensembl

Ensembl and UniProtKB are closely aligned in the production of the annotated genome. UniProtKB data is used as one of the inputs for the Ensembl gene build process.

After Ensembl completes a gene build, cross-referencing in UniProtKB is carried out on the basis of exact matching between UniProtKB isoforms and the Ensembl transcripts.

Current Status of Mapping

	Human	Mouse
Ensembl (Release 93)		
Genes (protein coding)	20,376	22,628
Transcripts (protein coding)	91,308	45,664
Transcripts with no UniProt Xref	922	28,639
UniProtKB (Release 2018_08)		
UniProtKB Entries in Proteome	20,381	16,987
Entries with Ensembl Transcript Xref	19,374	15,065
Entries with no Ensembl Xref	1,007	1,922

Cross-referencing in UniProtKB to human and mouse transcripts is extensive, but not complete. Where entries, or isoforms within entries, appear to be missing in UniProtKB, this is almost always because the relevant protein sequences are held in the unreviewed section of UniProtKB, and have not yet been incorporated into the manually reviewed entries where they belong.

Mapping Improvement

Discussion between UniProt and Ensembl on how to improve the extent and reliability of mappings has led to the creation of a common workspace in which the mapping between UniProtKB isoforms and Ensembl transcripts can be reviewed, agreed and updated.

Planned Outcomes

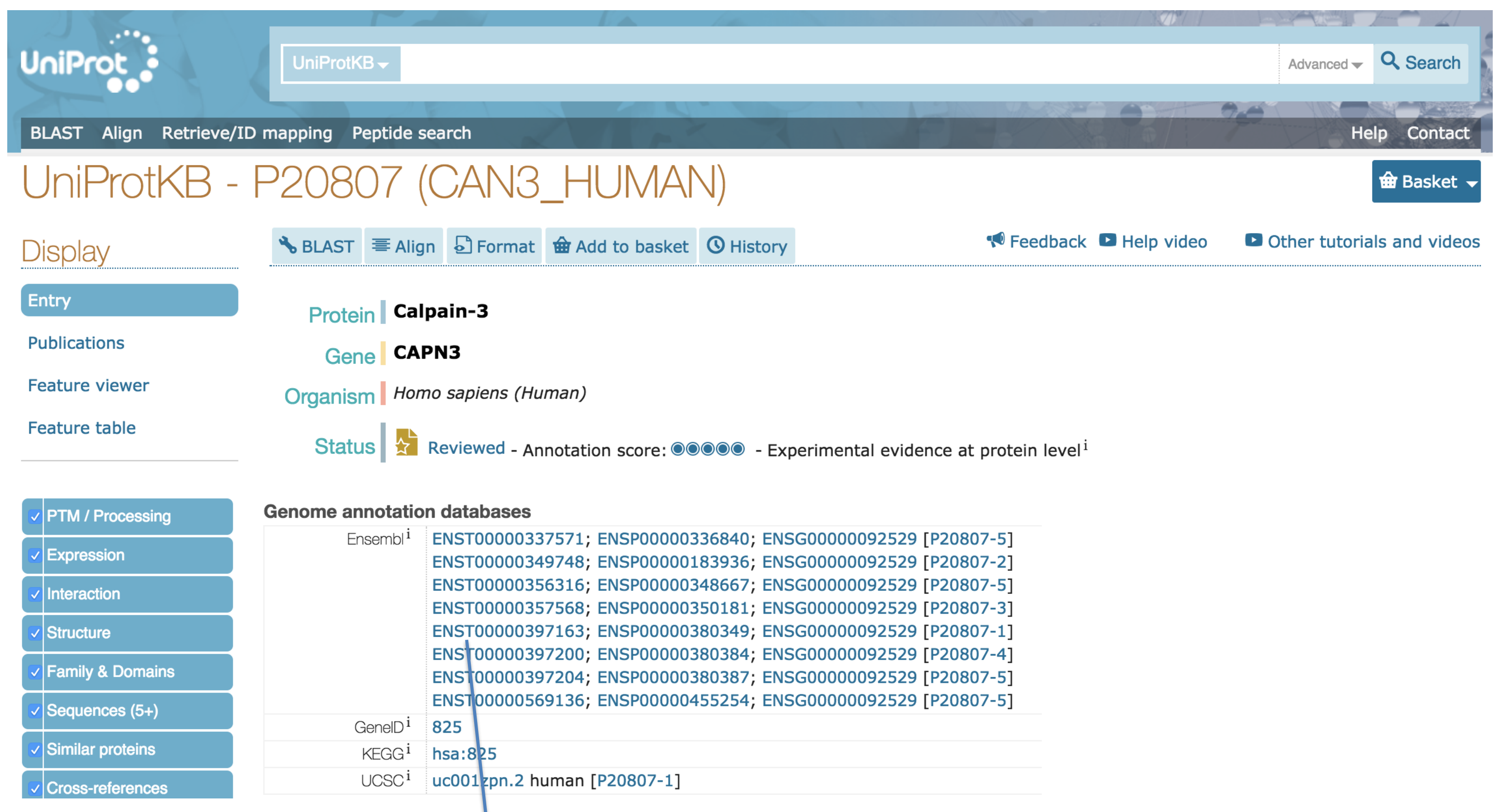
An important outcome of this joint UniProt : Ensembl collaboration will be a publicly available mapping database, which will provide protein : transcript mappings that remain stable through the different release cycles of UniProtKB and Ensembl.

The database will also provide representative sets of proteins and transcripts that are reliably mapped for different genomes, starting with human and extending to mouse and other organisms.

The new database will allow UniProtKB to include Ensembl cross-referencing with the proteome datasets that are currently provided.

Sequence Position Cross-referencing

Exact matching of sequence numbering plays a key role in making it possible to easily build datasets which combine genomic and protein data.

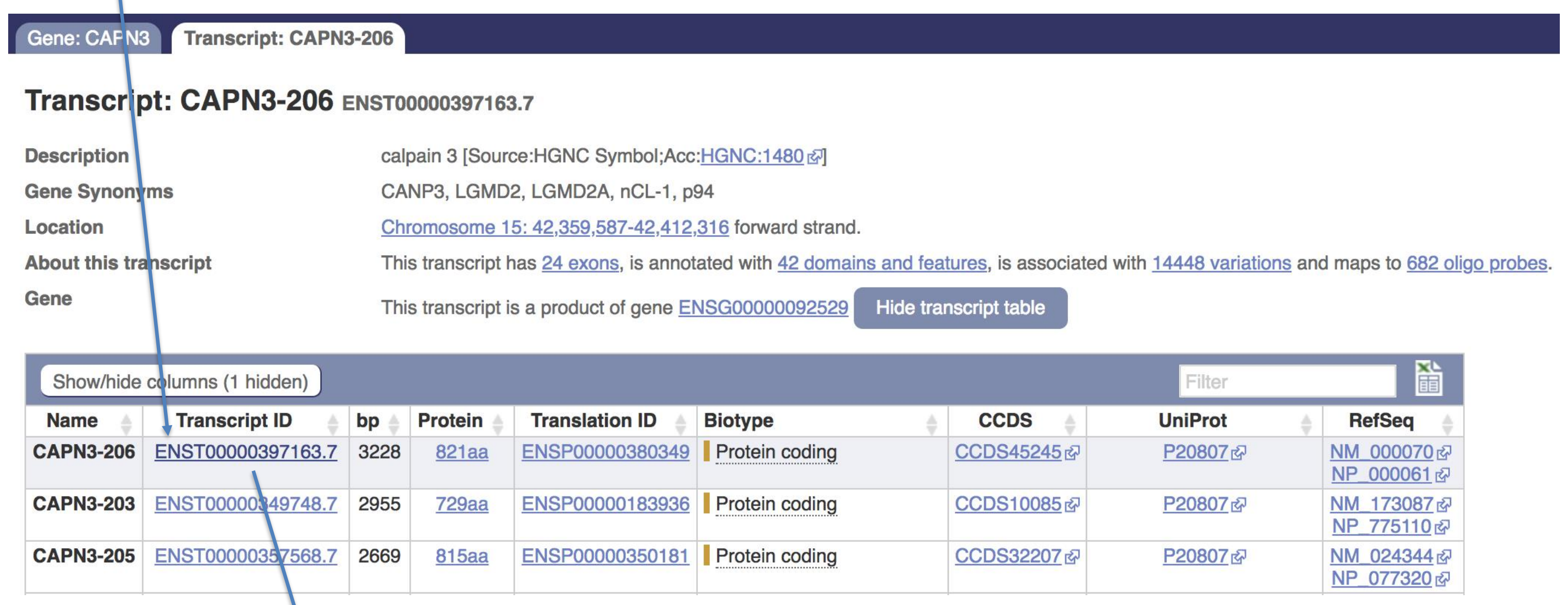


UniProtKB - P20807 (CAN3_HUMAN)

Protein | Calpain-3
Gene | CAPN3
Organism | *Homo sapiens (Human)*
Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

Genome annotation databases

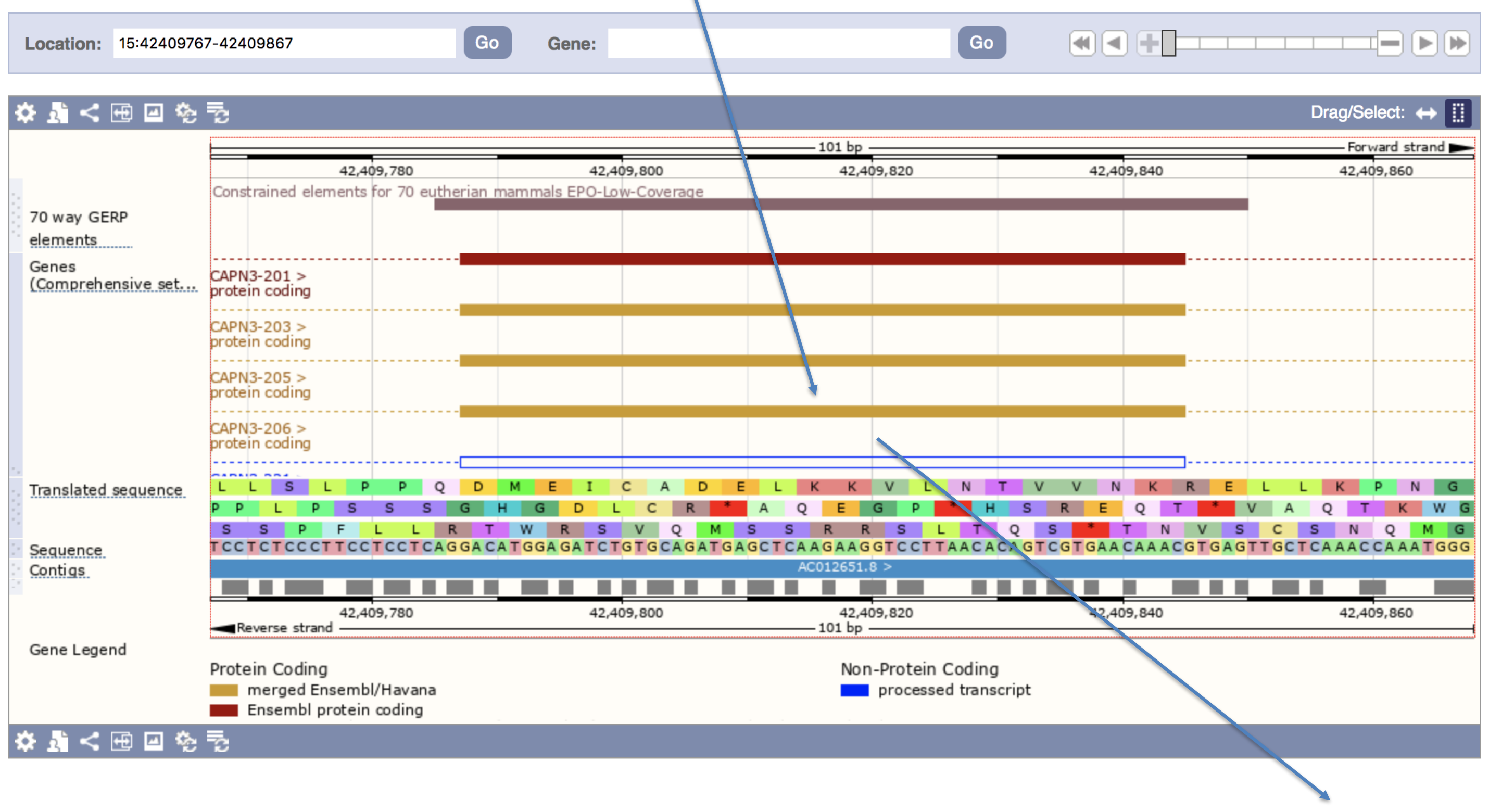
Ensembl	ENST00000337571; ENSP00000336840; ENSG00000092529 [P20807-5] ENST00000349748; ENSP00000183936; ENSG00000092529 [P20807-2] ENST00000356316; ENSP00000348667; ENSG00000092529 [P20807-5] ENST00000357568; ENSP00000350181; ENSG00000092529 [P20807-3] ENST00000397163; ENSP00000380349; ENSG00000092529 [P20807-1] ENST00000397200; ENSP00000380384; ENSG00000092529 [P20807-4] ENST00000397204; ENSP00000380387; ENSG00000092529 [P20807-5] ENST00000569136; ENSP00000455254; ENSG00000092529 [P20807-5]
GeneID	825
KEGG	hsa:825
UCSC	uc001zpn.2 human [P20807-1]



Transcript: CAPN3-206 ENST00000397163.7

Description: calpain 3 [Source:HGNC Symbol;Acc:HGNC:1480]
Gene Synonyms: CANP3, LGMD2, LGMD2A, nCL-1, p94
Location: Chromosome 15: 42,359,587-42,412,316 forward strand.
About this transcript: This transcript has 24 exons, is annotated with 42 domains and features, is associated with 14448 variations and maps to 682 oligo probes.
Gene: This transcript is a product of gene ENSG00000092529

Name	Transcript ID	bp	Protein	Translation ID	Biotype	CCDS	UniProt	RefSeq
CAPN3-206	ENST00000397163.7	3228	821aa	ENSP00000380349	Protein coding	CCDS45245	P20807	NM_000070 NP_000061
CAPN3-203	ENST00000349748.7	2955	729aa	ENSP00000183936	Protein coding	CCDS10085	P20807	NM_173087 NP_775110
CAPN3-205	ENST00000357568.7	2669	815aa	ENSP00000350181	Protein coding	CCDS32207	P20807	NM_024344 NP_077320

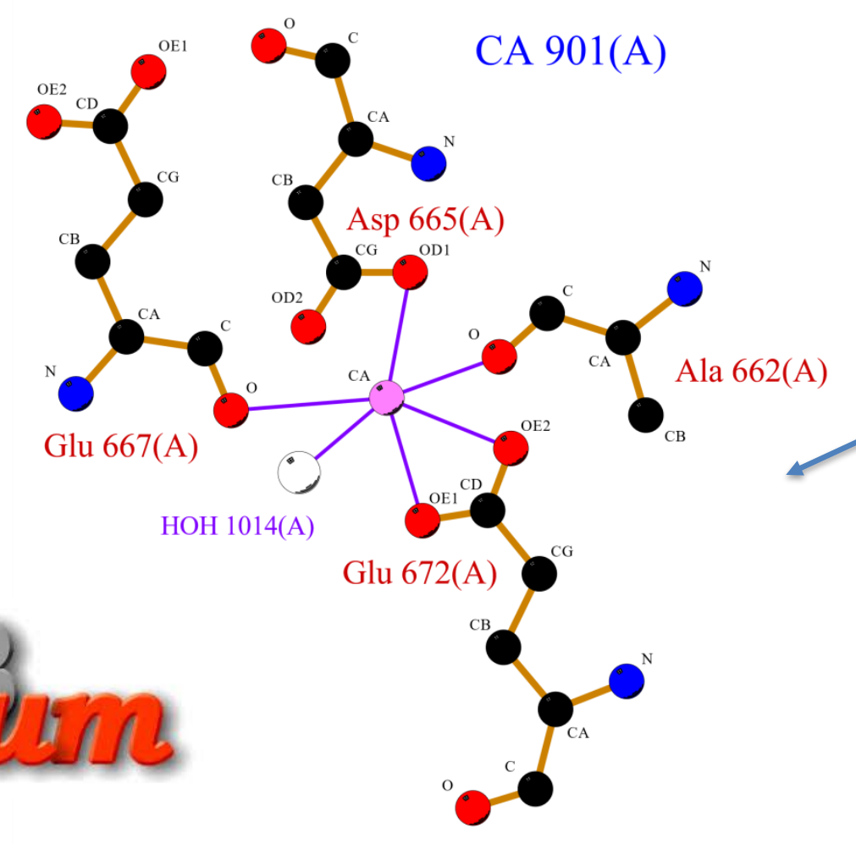


```
2161 AGTGAGGAACAGCAACAATTCCGGAACATTTTCAAGCAGATAGCAGGAGATGACATGGAG   2220
1942 AGTGAGGAACAGCAACAATTCCGGAACATTTTCAAGCAGATAGCAGGAGATGACATGGAG   2001
 648 -S--E--E--Q--Q--Q--F--R--N--I--F--K--Q--I--A--G--D--D--M--E-   667

2221 ATCTGTGCAGATGAGCTCAAGAAGGTCCTTAACACAGTCGTGAACAAACACAAGGACCTG   2280
2002 ATCTGTGCAGATGAGCTCAAGAAGGTCCTTAACACAGTCGTGAACAAACACAAGGACCTG   2061
 668 -I--C--A--D--E--L--K--K--V--L--N--T--V--V--N--K--H--K--D--L-   687

2281 AAGACACACGGGTTCACACTGGAGTCCTGCCGTAGCATGATTGCGCTCATGGATACAGAT   2340
2062 AAGACACACGGGTTCACACTGGAGTCCTGCCGTAGCATGATTGCGCTCATGGATACAGAT   2121
 688 -K--T--H--G--F--T--L--E--S--C--R--S--M--I--A--L--M--D--T--D-   707
```

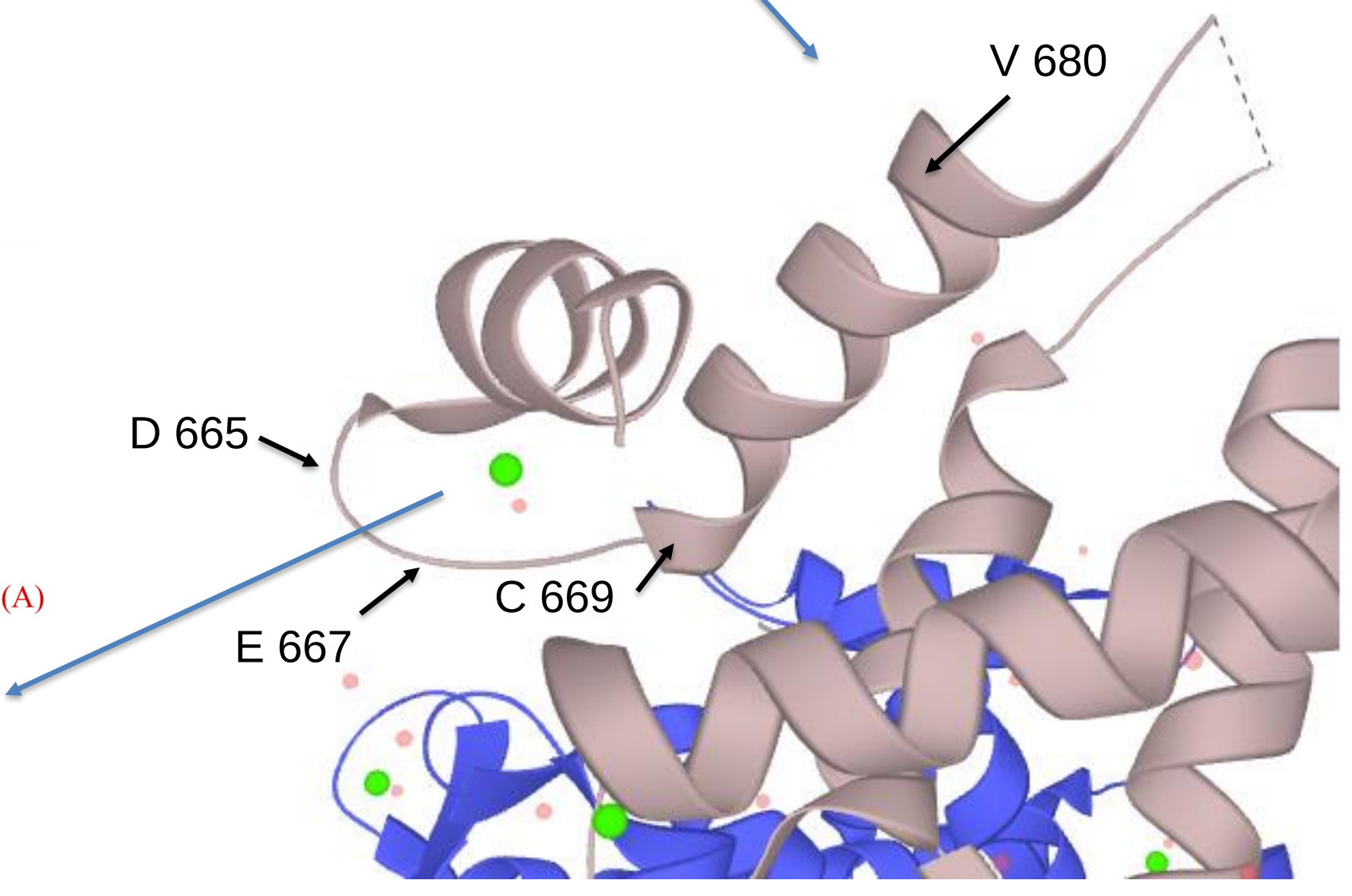
cDNA

```
       610        620        630        640        650
PIIFVSDRAN SNKELGVDQE SEEGKGKTSP DKQKQSPQPQ PGSSDQESEE
       660        670        680        690        700
QQQFRNIFKQ IAGDDMEICA DELKKVLNTV VNKHKDLKTH GFTLESCRSM
       710        720        730        740        750
IALMDTDGSG KLNLQEFHHL WNKIKAWQKI FKHYDTDQSG TINSYEMRNA
       760        770        780        790        800
VNDAGFHLNN QLYDIITMRY ADKHMNIDFD SFICCFVRLE GMFRAFHAFD
       810        820
KDGDGIIKLN VLEWLQLTMY A
```

protein sequence



PDBsum



PDBe

UniProt e! EMBL-EBI PIR SIB Swiss Institute of Bioinformatics

European Bioinformatics Institute
Wellcome Genome Campus
Hinxton, Cambridge, CB10 1SD, UK

Tel. +44 (0) 1223 494 444
comms@ebi.ac.uk
www.ebi.ac.uk