Gene Assignment: Gm7609

Elizabeth Nguyen

**Introduction**

Gene clusters may include large repeats of duplicated genes. The *Sp100-rs* repeat cluster is a long-range structure found on chromosome 1 of the house mouse, *Mus musculus* [1]. Long repeat regions are extremely difficult to sequence and result in gaps in “complete” genome sequences [2]. The *Sp100-rs* gene is a chimeric fusion of *Sp100* and *Csprs*, two other genes associated with the *Sp100-rs* cluster [1]. The *predicted pseudogene 7609*, or *Gm7609*, is one of many variable copies of the *Sp100-rs* gene resulting from duplication. NCBI classifies this *Gm7609* as a protein coding gene because its coding sequence appears to be intact. However, it is unclear whether this gene is transcribed *in vivo*. Annotating *Gm7609* will help to close a gap in the *Mus musculus* genome.

**Methods**

I first searched for Gm7609 using the NCBI gene database, the Ensembl genome browser, and the USCB genome browser. For further analyses, I choose to use the NCBI gene accession number NC\_000067.6 and the curated protein associated with *Gm7609* provided by NCBI under NP\_001075215.1.

**Results**

**Structure:** I first wanted to structurally annotate this gene. According to the NCBI gene page (accessed 3/4/2020; updated 1/23/2020), this 6 exon gene has 5 transcripts. However, upon examining the transcripts sequenced in NCBI *Mus musculus* Annotation Release 108 (8/10/2019), they appear to only have 3-5 exons each (Figure 1A). In contrast to the NCBI result(s), a search for *Gm7609* using Ensembl (accessed 3/4/2010; release 99, 1/16/2020) only associates it with two transcripts, *Gm7609-201* and *Gm7609-202*, with 4 and 3 exons respectively (Figure 1B). The UCSB Genome Browser (GRCm38/mm10) displays 4 transcripts with 2-4 exons (Figure 1C). These discrepancies between major genome browsers speak to the poor annotation regarding this gene.

**Function:** In order to identify the putative function of the protein associated with *Gm7609*, I submitted the protein query to the National Library of Medicine’s Conserved Domain Database [3] through NCBI. Only one conserved domain was found between position 106-235 (E-value 1.24e-04): 7tm\_1 super family (accession # cl37946) detected with pfam00001. This 7 transmembrane receptor of the rhodopsin family contains G-protein coupled receptors, including opsins. Running the protein query through blastp (protein-protein blast) against the Non-redundant protein sequences (nr) database [4] identified numerous proteins with high query coverage and sequence identity. As expected, many of these proteins were G-protein coupled receptors or Sp100-rs or Csprs protein-related. Submitting the FASTA file for NP\_001075215.1 to Interpro sequence analyses [5] also found the 7 transmembrane GPCR rhodopsin-like domain (IPR017452). Submitting *Gm7609* to AmiGO2 powered by PANTHER produced no significant results for biological process, molecular function, or cellular component. Submitting *Gm7609* to Reactome or Kyto Encyclopedia of Genes and Genomes produced no results.

In order to assess evolutionary constraint, I ran the NP\_001075215.1 through blastp using the Non-redundant protein sequences (nr) database. This search found high conservation for this protein across a number of rodent species, as well as primates, carnivores, odd-toed ungulates, and even-toed ungulates (Figure 2). This suggests that the resultant protein of *Gm7609* serves an important biological role. According to the NCBI gene page (accessed 3/4/2020; updated 1/23/2020), homologs of the Gm7609 gene can be found in wolves (*C. lupus*), cows (*B. Taurus*), and rats (*R. norvegicus*).

**Regulation:** In order to explore gene regulation, I assessed the regulatory build in Ensembl [6] (Figure 3A). This gene is associated with a promoter, two transcription factor binding sites, and one promoter flanking region. The promoter region is poised (has the potential to be activated) in the forebrain and midbrain of E10.5 mice. However, the promoter for this gene is typically inactive (bears no epigenetic modifications) or epigenetically repressed. Accordingly, there does not seem to be any open chromatin structures present in this gene. This explains why even though *Gm6709*’s coding sequence appears to be intact, it is unclear whether this gene is transcribed *in vivo*. Neither associated transcription factor binding sites display an active or potentially active epigenetic signature. Proximal enhancers defined as the promoter flanking region [7], however, show an active epigenetic signature in a C57BL/6 embryonic cell line and E14.5-15 forebrain, hindbrain, and midbrain. Additionally, this promoter flanking region has the potential to be activated in the midbrain of E13.5 mice. In support of the Ensembl findings, the UCSB Genome Browser [8] shows heterochromatin H3K9me3-associated (Hc-H) expression in the forebrain and heart of E11 and E15 mice respectively (Figure 3B). This histone mark is associated with gene repression [9].

**Orthology:** To identify orthologs for the curated protein associated with *Gm7609*, I ran NP\_001075215.1 through blastp using the Model Organisms (landmark) database and the PSI-BLAST algorithm [10]. This produced BLAST hits only in *Mus musculus* and *Danio rerio* (bony fish). The top result for *Danio rerio* is olfactory receptor 2T2-like (XP\_009296720.1). However, while it had a relatively low E-value (5e-37) with 95% query coverage, the percent identity was only 34.15%. I then refined the precision matrix with a second iteration selecting for only E-values less than 2e-12. This identified more *Mus musculus* and *Danio rerio* proteins, as well as some *Homo sapiens* (humans) and *Drosophila melanogaster* (flies) proteins. The top hit for *Homo sapiens* was probable G-protein coupled receptor 148 (NP\_997247.2). This protein had 72% query coverage and an E-value of 3e-14, but only 18.86% identity. The top hit for *Drosophila melanogaster* was leucokinin receptor (NP\_647968.3). This protein had 89% query coverage and an E-value of 9e-09, but only 15.49% identity. These numbers suggest unlikely cross-species homology. To confirm this hypothesis, I performed a reciprocal blastp search against the Non-redundant protein sequences (nr) database for *Mus musculus* using the top hits in each of these three species. As expected, no significant similarity was found for any of these proteins using these criteria. When I searched for this gene in Genomicus to compare syntenic regions between species, *Gm7609* and ENSMUSG00000079457 did not appear to match any genes indicating no homologues in humans.

**Localization:** According to NCBI, the protein sequence for *Gm7609* is the same as that of *Csprs*. Accordingly, a UniProKB search [11] for *Gm7609* brings up entry Q3U0D6, an uncharacterized protein associated with 3 genes: *Csprs*, *Gm7592*, and *Gm7609*. Sourced via Interpro and UniProKB-KW, this protein is associated with 3 GO terms: membrane, integral component of membrane, and nucleus. Sequence analyses identified six transmembrane segments as denoted by α-helical transmembrane regions and β-barrel transmembrane proteins. This suggests that the *Gm7609*-associated protein is a transmembrane protein located on the nuclear envelope (Figure 4). The hydrophobicity profile for the uncharacterized mouse protein Q3U0D6 sequence predicted using the Kyte-Doolittle [12] scale via Expasy protscale algorithm supports this protein’s transmembrane status (Figure 5).

**Tissues of expression:** According to the mouse ENCODE transcriptome data displayed on the NCBI page for this gene, *Gm7609* shows highest expression levels in the adult spleen (1.741 RPKM), thymus (1.434 RPKM), and mammary glands (0.981 RPKM) [13]. It is expressed at lower levels in a number of other tissue, with moderate expression in the large intestine, bladder, heart, kidney, subcutaneous adipose tissue (Figure 6).

**Discussion**

*Gm7609* may code for a transmembrane protein at the level of the nuclear membrane. It is expressed in a number of tissue with the highest level of expression in the adult spleen, thymus, and mammary glands. However, as referenced in previous findings, it is unclear whether or not this gene is transcribed *in vivo*. Findings in this paper regarding the regulation of the gene support this assertion. For example, the promoter associated with *Gm7609* only has the potential to be activated during the embryonic stage; even so, *Gm7609* expresses a histone mark associated with gene repression during that stage. Despite the evolutionary constraint placed on the curated protein associated with *Gm7609*, the orthology-related findings are ambiguous. More research is needed to better understand the biological significance of this gene and its associated products, if any.

**References**

[1] W. Traut, I. M. Rahn, H. Winking, B. Kunze, and D. Weichenhan, “Evolution of a 6–200 Mb long-range repeat cluster in the genus Mus,” *Chromosoma*, vol. 110, no. 4, pp. 247–252, Aug. 2001, doi: 10.1007/s004120100152.

[2] T. J. Treangen and S. L. Salzberg, “Repetitive DNA and next-generation sequencing: computational challenges and solutions,” *Nat. Rev. Genet.*, vol. 13, no. 1, pp. 36–46, Jan. 2012, doi: 10.1038/nrg3117.

[3] S. Lu *et al.*, “CDD/SPARCLE: the conserved domain database in 2020,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D265–D268, Jan. 2020, doi: 10.1093/nar/gkz991.

[4] “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Res.*, vol. 46, no. Database issue, pp. D8–D13, Jan. 2018, doi: 10.1093/nar/gkx1095.

[5] A. L. Mitchell *et al.*, “InterPro in 2019: improving coverage, classification and access to protein sequence annotations,” *Nucleic Acids Res.*, vol. 47, no. Database issue, pp. D351–D360, Jan. 2019, doi: 10.1093/nar/gky1100.

[6] F. Cunningham *et al.*, “Ensembl 2019,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D745–D751, Jan. 2019, doi: 10.1093/nar/gky1113.

[7] D. R. Zerbino, S. P. Wilder, N. Johnson, T. Juettemann, and P. R. Flicek, “The Ensembl Regulatory Build,” *Genome Biol.*, vol. 16, no. 1, 2015, doi: 10.1186/s13059-015-0621-5.

[8] W. J. Kent *et al.*, “The Human Genome Browser at UCSC,” *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jun. 2002, doi: 10.1101/gr.229102.

[9] J. S. Becker, D. Nicetto, and K. S. Zaret, “H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes,” *Trends Genet. TIG*, vol. 32, no. 1, pp. 29–41, Jan. 2016, doi: 10.1016/j.tig.2015.11.001.

[10] S. F. Altschul *et al.*, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

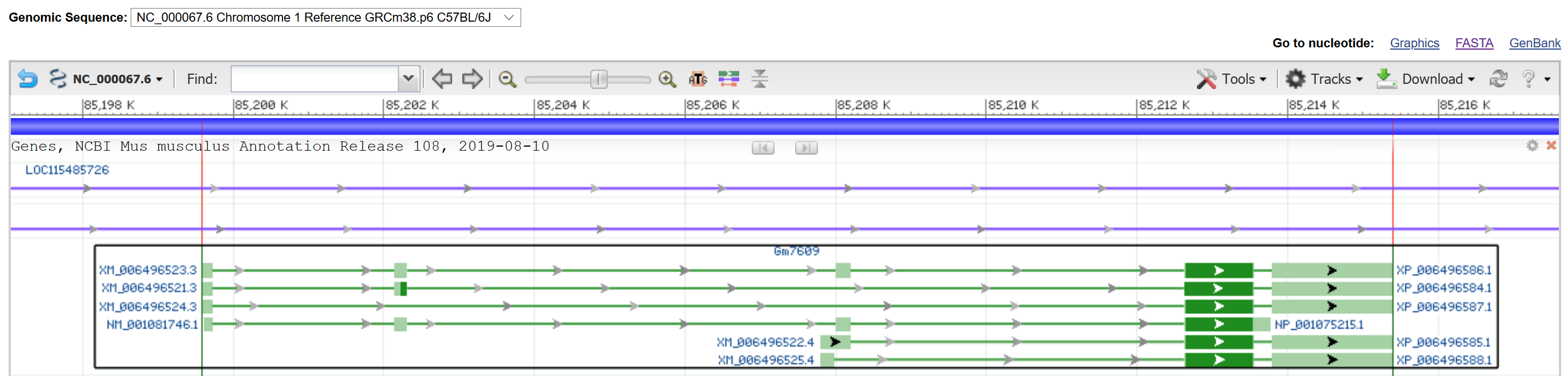
[11] “UniProt: a worldwide hub of protein knowledge,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.

[12] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydropathic character of a protein,” *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–132, May 1982, doi: 10.1016/0022-2836(82)90515-0.

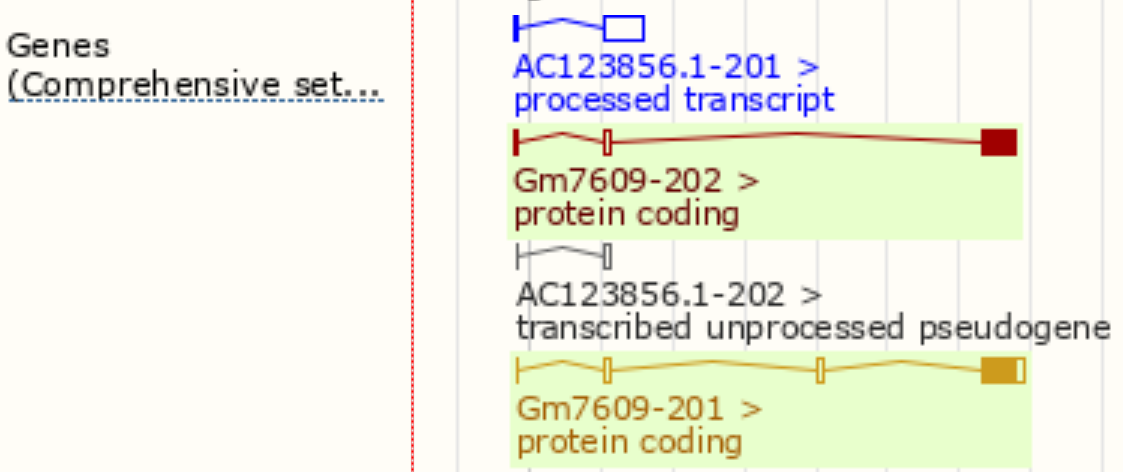
[13] F. Yue *et al.*, “A Comparative Encyclopedia of DNA Elements in the Mouse Genome,” *Nature*, vol. 515, no. 7527, pp. 355–364, Nov. 2014, doi: 10.1038/nature13992.

Figure 1. Transcripts and exons of *Gm7609*

1. *Gm7609* as shown using NCBI *Mus musculus* Annotation Release 108



1. *Gm7609* as shown using Ensembl Release 99



1. *Gm7609* as shown using UCSB Genome Browser (GRCm38/mm10)

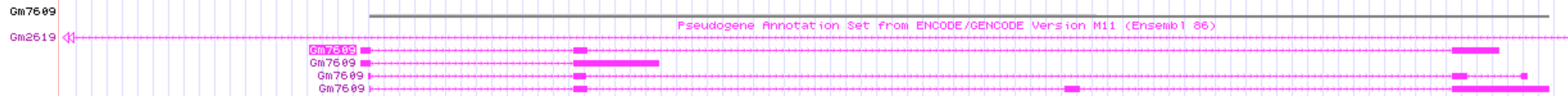
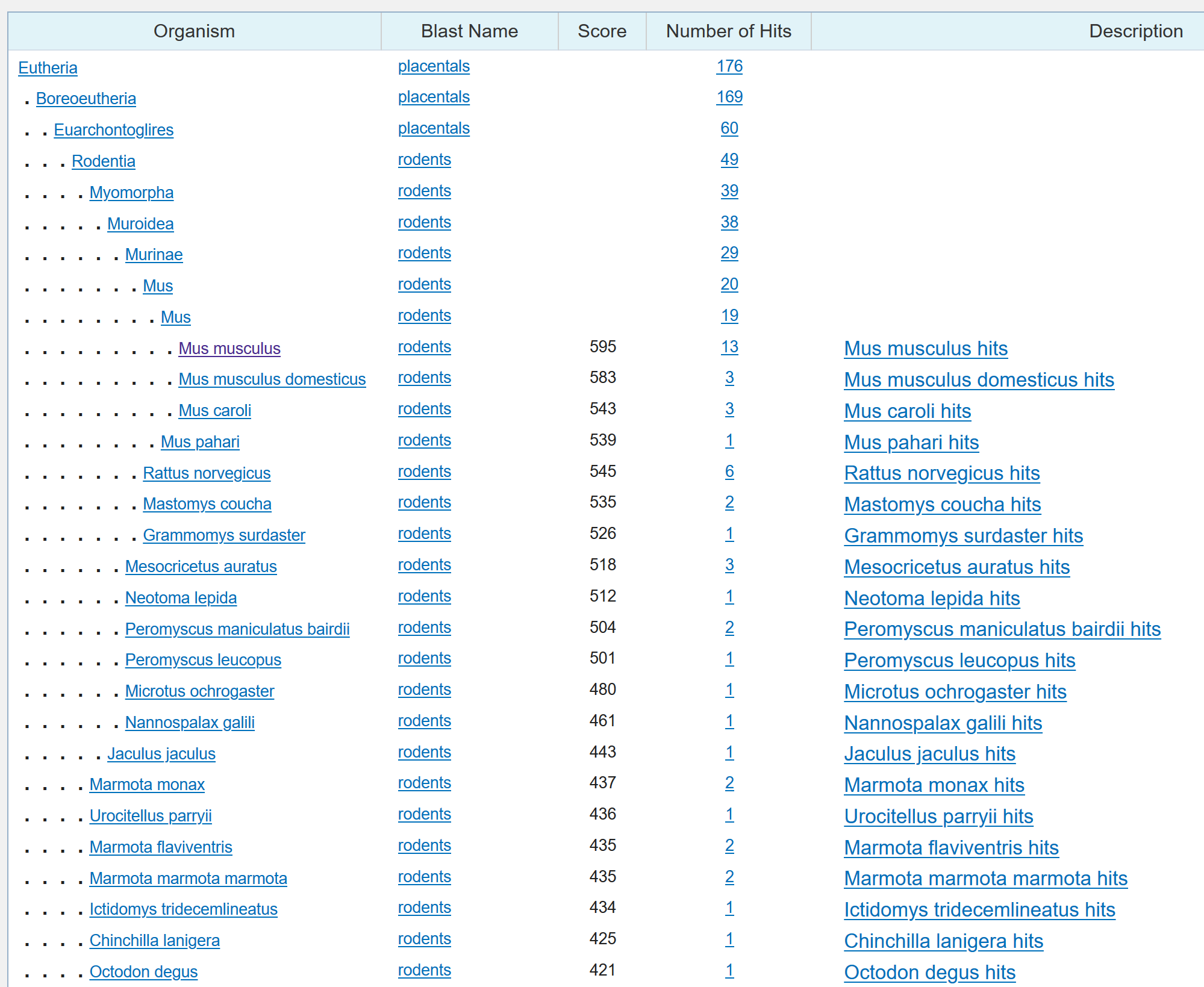


Figure 2. Lineage of organisms with significant sequence alignments for NP\_001075215.1



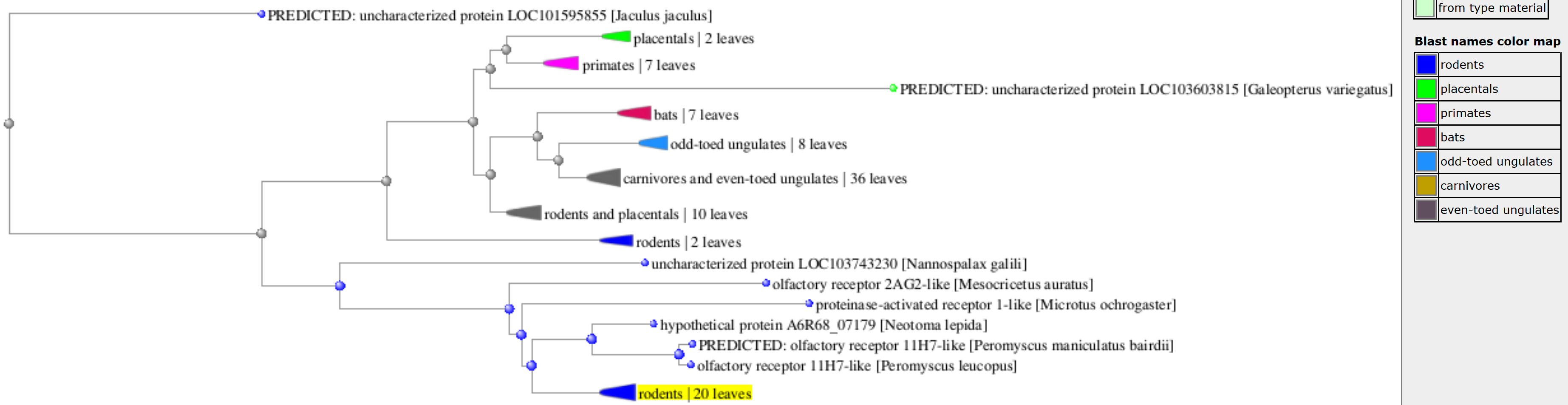
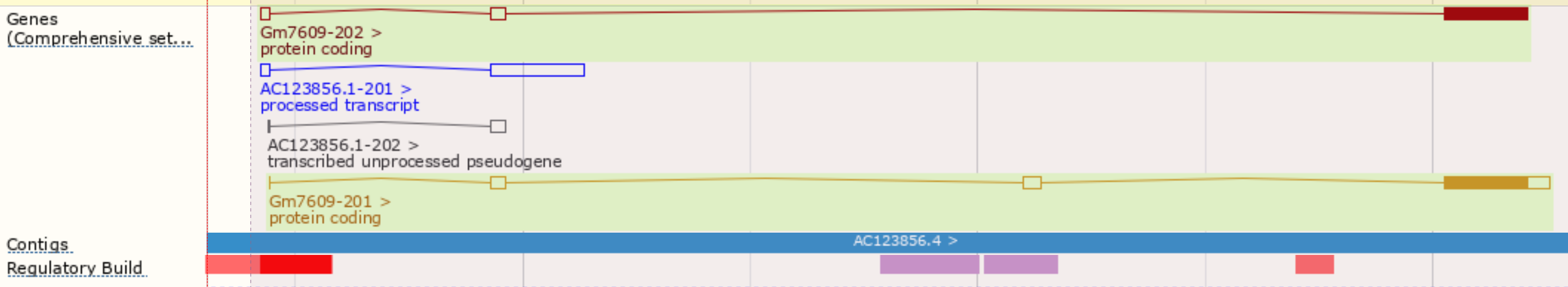


Figure 3: Gene regulation associated with *Gm7609*

1. Regulatory build in Ensembl





1. UCSB Genome Browser chromatin state from embryonic tissue

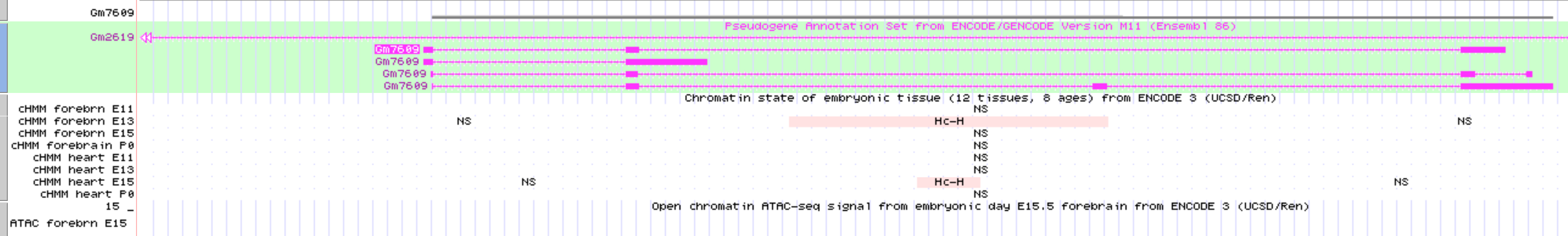


Figure 4. Subcellular localization of *Gm7609*-associated protein

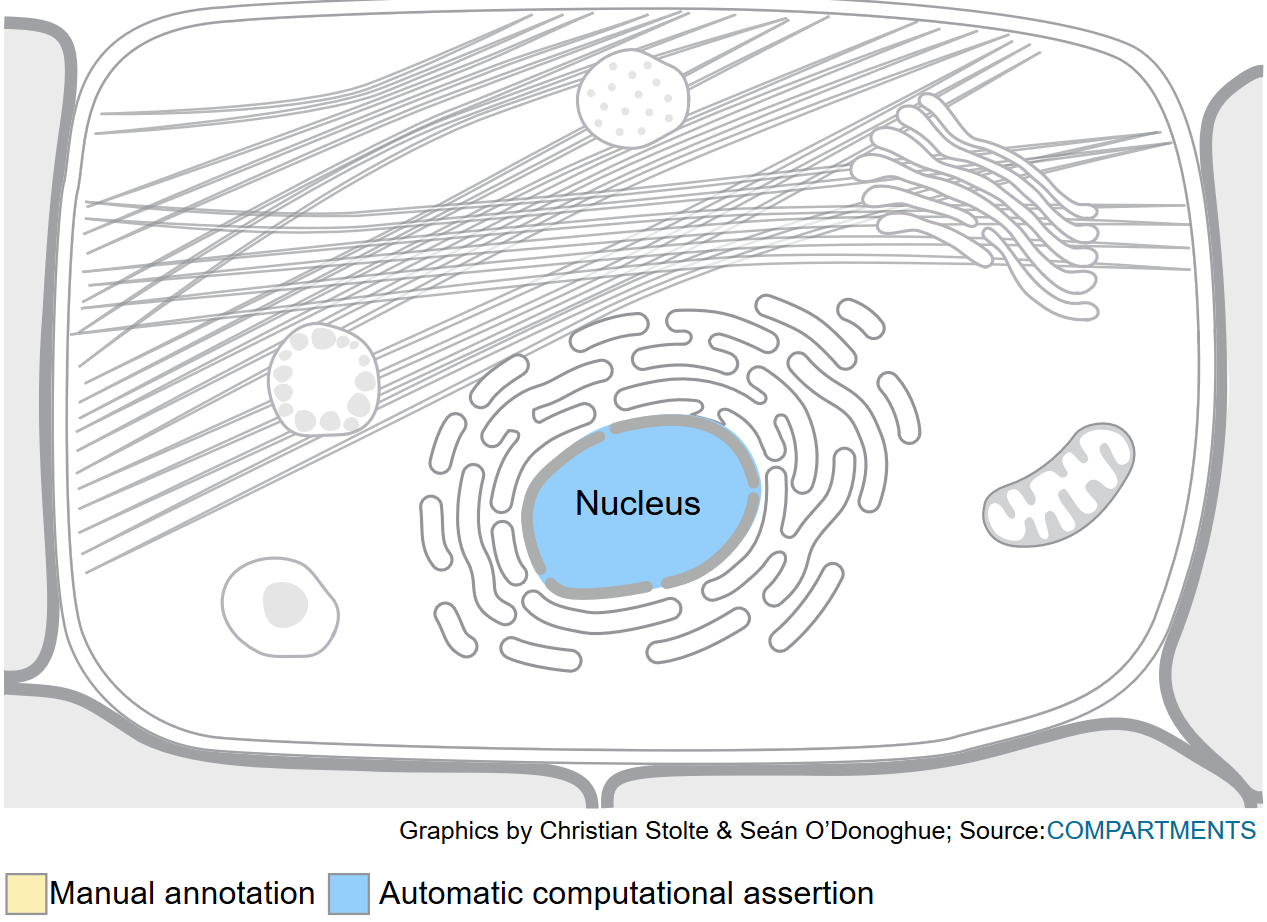


Figure 5: Hydrophobicity profile for uncharacterized mouse protein Q3U0D6

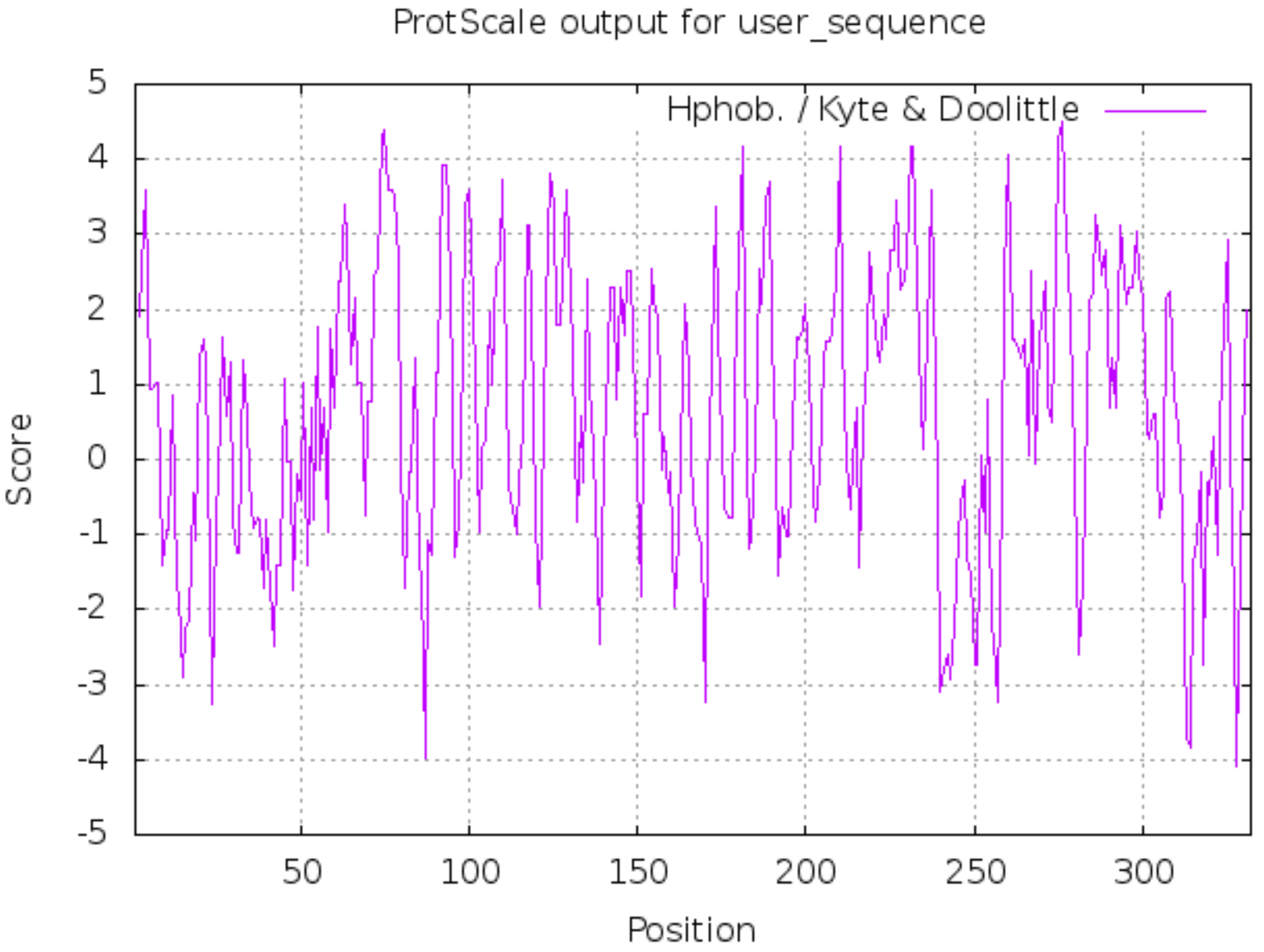
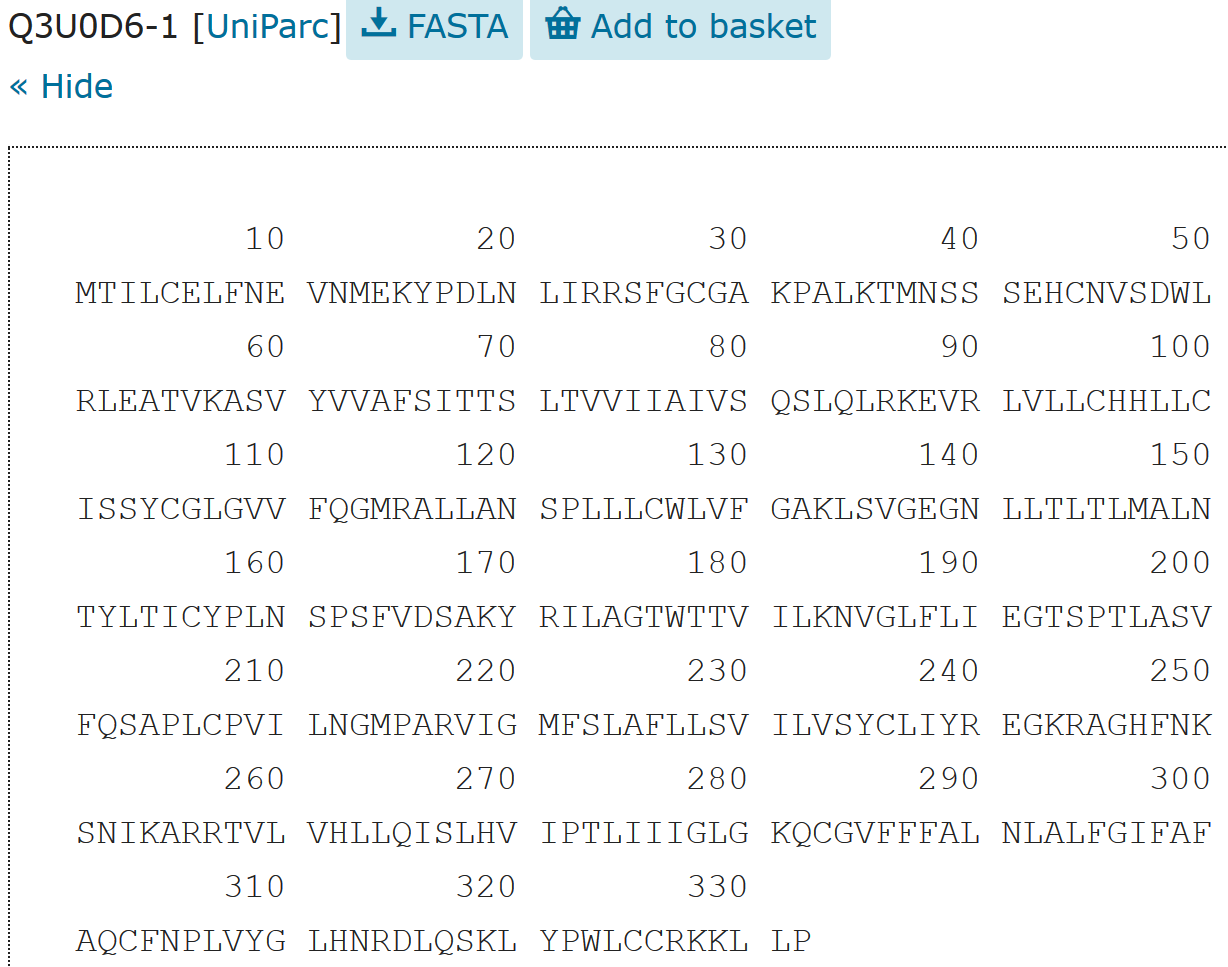


Figure 6. Mouse ENCODE transcriptome data for *Gm7609*

