William McElhenney

*Initial Sequence Evaluation*

The first step performed was BLAST and FASTA searches with the three given sequences (gene, mRNA, and promoter). BLAST searches were ran against the Mammalian taxon with otherwise default settings. FASTA searches were ran against the Mouse and Human databases as running against the full Mammalian database produced results that, while they aligned well, did not tell us much about the sequence.

For the gene sequence, BLAST returned a high similarity to *Mus musculus* Fcamr gene for Fca/m receptor (ACC: AB071978.1) at 99.66% identity, an E-value of 0.0, and query coverage of 70%. The other results from this search corresponded to the same gene. The default FASTA search returned similar results with Fcamr being the first in the list with lowest E-value.

For the mRNA sequence, BLAST and FASTA searches returned high similarity to the same *Mus musculus* Fcamr gene as the gene sequence did.

The promoter sequence provided a few new and relevant details. The BLAST and FASTA searches for this sequence indicated high similarity to the *Mus musculus* butyrophilin (Btn1a1) gene, *promoter region,* and complete cds (ACC: U67065.1). Intriguingly, it can be seen in these results that the bp region 510 – 605 seems to have high repetition in the mouse genome aligning well multiple chromosomes (average E-value is approximately 1e-26 for this portion of the promoter), which indicates a shared signaling motif within this promoter (we did not try to identify it). Butyrophilin is a nice result, as Entrez states that “Butyrophilin is the major protein associated with fat droplets in the *milk* . . . and may have arisen relatively recently in evolution by the *shuffling of exons* between 2 ancestral gene families”.1

*Exon Sequence Evaluation*

Having completed the naive approach to identifying these sequences, the exons from the gene were then run against BLAST and FASTA. For the BLAST settings, the word size was reduced to 16, and the filter for low-complexity regions was turned off. FASTA settings were left as before in *Initial Evaluation*.

Exon 1 (bp 1 – 510) aligned well with the already identified Fcamr gene in *Mus musculus* in both BLAST and FASTA.

Exons 2 and 3 (bp 1401 – 1640 and bp 2299 – 2538) are the same sequence. This sequence aligned with a new gene in BLAST, *Homo sapiens* gene for immunoglobulin heavy chain variable region (ACC: AB203310.1), with an E-value of 0.042, 96% identity, and 11% query coverage. This is not a particularly high quality alignment, but it is interesting in relation to Butyrophilin, as Butyrophilin is a member of the Immunoglobulin superfamily.1 These results were mirrored in FASTA but with much better statistics (E-values were below 1e-10 and query coverage was greater than 50%).

Exon 4 (bp 2941 – 3081) produced 1 very poor result in BLAST, but FASTA produced several results indicating a likeness to Butyrophilin. These results further implicate this gene as having a similar structure/function to Butyrophilin.

Exon 5 (bp 3672 – 5121) returned very good likeness to the Fcamr gene in both BLAST and FASTA.

*Protein Structure and Function Hypothesis*

Based on the results of the above sequence testing, we propose that the provided sequences correspond to a Butyrophilin-like protein produced by the human mammary glands (particularly similar to Btn1a1, but we will be referring to this as Butyrophilin). This protein likely has immunoglobulin like regions with Fcamr like receptor regions. This structural hypothesis is in line with what the literature states about about Butryophilin in that Butryophilin is a member of the immunoglobulin superfamily and “may act as a specific membrane-associated receptor.”2 Based on this hypothesis we further hypothesize that this protein was produced from the exon shuffling of an Fcamr gene and a Butryophilin gene.

Functionally, we expect that this protein is a membrane protein associated with the production of fat-droplets in milk, as Butryophilin is known to be.

*Protein Structure and Function Evaluation*

The mRNA was translated using the ExPASy translate tool. In 5’3’ Frame 1 a 493 residue sequence was found. The length of this sequence is comparable to human and mouse Butryophilin (available data suggests both range from about 450 to 530 amino acids in length), so we believe this to be a good choice.

This sequence was then ran through BLASTp and FASTA to determine if this sequence was homologous to both Butryophilin and Fcamr, as we expect it to be for the exon shuffling theory.

References:

1. <https://www.ncbi.nlm.nih.gov/gene?cmd=Retrieve&dopt=full_report&list_uids=696>

2. <https://www.uniprot.org/uniprot/Q13410#function>

3. https://www.omim.org/entry/601610

4.