Published in final edited form as:

Protein Expr Purif. 2007 August ; 54(2): 275–282. doi:10.1016/j.pep.2007.03.017.

Expression of Recombinant Protein Encoded by *LOC387715* in *Escherichia coli*

Dequan Chen[1,*], **Marlyn P. Langford**[2], **Chris Duggan**[2], **Benjamin J. Madden**[3], and **Albert O. Edwards**[1]

[1] Institute for Retina Research, Presbyterian Hospital 8210 Walnut Hill Lane, Dallas, TX 75231 and Departments of Ophthalmology, Mayo Clinic College of Medicine, 200 1st ST SW, Rochester, MN 55905

[2] Department of Ophthalmology, Louisiana State University Health Sciences Center, 1501 Kings Highway, Shreveport, LA 71130

[3] Mayo Proteomics Research Center, Mayo Clinic College of Medicine, 200 1st ST SW, Rochester, MN 55905

Abstract

LOC387715 is a hypothetical gene located on human chromosome 10q26.13 that is associated with the development of age-related macula degeneration (AMD). The native open reading frame (ORF) of *LOC387715* cDNA – LOC387715(ORF), contains a large number of *Escherichia coli* (*E. coli*) rare codons (RC) including 5.6% and 15.0% Group-I and IIa translational problem causative (TPC) RCs respectively, which forms 3 and 4 simple *E. coli* rare codon clusters (RCC) where RCs are spaced by 1 and 2 respective non-TPC codons and one complex *E. coli* RCC where RCs and RCCs are spaced by < 5 non-TPC codons. We modified the entire 35 *E. coli* RCs (6, 16 and 13 Group I, IIa and IIb RCs respectively) present in LOC387715(ORF) with their optimal or sub-optimal synonymous degenerate codons, and the resulted LOC387715(ORF)m was free from Shine-Dalgarno-like sequence (SDLS) and ribosome binding site complementary sequence (RBSCS). SDS-PAGE and Western blotting analysis demonstrated that LOC387715(ORF)m was capable of highly expressing the recombinant protein rLOC387715 in *E. coli*. Mass spectrometry analysis indicated that the bacterial expressed rLOC387715 contained the correct and expected amino acid (aa) sequence without aa misincorporation, aa missing or frame-shift. The results suggest that high and authentic expression of *LOC387715* recombinant protein in *E. coli* was achieved by the synonymous modification of its native ORF cDNA sequence for all the 3 groups of bacterial RCs and the simultaneous elimination of SDLS and RBSCS sequences.

Keywords

LOC387715; age-related macular degeneration (AMD); single nucleotide polymorphism (SNP); nonsynonymous coding; synonymous modification; recombinant protein; rare codon; Shine-Dalgarno-like sequence; ribosome binding site complementary sequence; nano-flow liquid chromatography electrospray tandem mass spectrometry

[*] To whom correspondence should be addressed. Dequan Chen, Ph.D., Department of Ophthalmology, Mayo Clinic College of Medicine, Guggenheim Bldg, Room Gu 9-93, 200 1st ST SW, Rochester, MN 55905, Phone: 507-284-2199, Fax: 507-284-8566, chen.dequan@mayo.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

LOC387715 is a hypothetical gene that was originally found by the Gnomon gene prediction program during the annotation of human genome sequences. It was recently named *ARMS2* (*age-related maculopathy susceptibility 2*) by HUGO (the human genome organization) Gene Nomenclature Committee (HGNC). At present, it retains the status of the "hypothetical" in the NCBI (National Center for Biotechnology Information) database due to the reason that no protein(s) encoded by this "gene" has been reported yet. However, NIH MGC (The National Institutes of Health Mammalian Gene Collection) project 42 had obtained 2 cDNA clones from pre-eclamptic placental tissue (NCBI Accession #: BC066349 and BC090924) for this gene, suggesting that it is most likely a protein-encoding gene.

Chromosome 10q26 region was linked to the risk of AMD by early family-based genome-wide scan studies 8,15,19,21,35,45,46. In this chromosomal region, genetic studies have demonstrated the association of AMD with single nucleotide polymorphisms (SNP) of *LOC387715* locus and adjacent genes. The nonsynonymous coding polymorphism rs10490924 (encoding Ala69Ser) in the hypothetical gene (Fig. 1A) was suggested as the variation most likely to explain the association of the chromosomal region with AMD16,29,34,38. Moreover, some studies have shown that either chromosome 10q26.13 or rs10490924 of *LOC387715* strongly interacts with smoking during AMD development 33,46. To elucidate the mechanisms regarding how *LOC387715* locus and its adjacent genes are involved in AMD development, it will be very helpful to first determine whether *LOC387715* is a real gene that encodes a functional protein(s) and what the normal function of the encoded protein(s) is in human cells. The ORFs of the *LOC387715* predicted mRNA and the 2 NIH MGC project cDNA clones, all encode a "hypothetical" protein consisting of 107 amino acids (aa), but no proteins known up to now have significant homology with it. This brings about the difficulty in finding and determining the possible native protein(s) encoded by this gene as well as the normal function (s) for the protein(s) that may be identified in the future.

In this study, we synonymously modified the ORF cDNA sequence (BC066349) of *LOC387715* and highly expressed the recombinant protein (rLOC387715) encoded by the modified cDNA in bacterium *E. coli*. The bacterial expressed recombinant protein had the same aa sequence as that of the predicted protein, suggesting that it can be used for future investigation related to *LOC387715* such as determination of its possible native protein(s).

Materials and Methods

Bacteria, vector, enzymes and antibodies

One Shot® TOP10 *Escherichia coli (E. coli)* competent cells (for plasmid amplification) were purchased from Invitrogen (Carlsbad, CA). *E. coli BL21*, *BL21(DE3)* and *Rosetta 2(DE3)* competent cells were obtained from Novagen (Madison, WI). Expression vector pGS21a and mouse anti His-tag monoclonal antibody (mAb) was from GenScript Corporation (Piscataway, NJ). Restriction enzymes *Eco*RI and *Xho*I were from Boehringer-Mannheim (GmbH, Germany). Rabbit anti-GST polyclonal antibody was made as previously described 3. Porcine trypsin was obtained from Promega Corporation (Madison, WI). HRP-conjugated secondary goat anti-rabbit IgG (F(ab′)$_2$ fragment-specific) and donkey anti-mouse IgG (H+L) were purchased from Jackson ImmunoResearch Laboratories, Inc. (West Grove, PA).

LOC387715 cDNA synthesis, cloning and sequencing

Self-designed synonymously modified ORF cDNA sequence of *LOC387715* – LOC387715 (ORF)m was synthesized and cloned in-frame into the bacterial expression vector pGS21a between *EcoR*I and *Xho*I sites by GenScript Corporation. The insert was further confirmed using DNA sequencing.

SDS-PAGE and protein relative amount quantitation

SDS-PAGE analysis was performed as previously described 2. Digital images and protein relative amount of the total protein were obtained using HP ScanJet 6300C and the ImageJ software 5.

Western blotting analysis

Bacterial cell lysates were separated by SDS-PAGE and transblotted onto Hybond-P PVDF membranes (Amersham Biosciences, Little Chalfont, Buckinghamshire, England). After blots were probed with anti-GST or anti-His tag antibody and corresponding HRP-conjugated secondary antibody, metal enhanced DAB substrate kit (Pierce Biotechnology, Rockford, IL) was used to identify the expressed *LOC387715*-encoded recombinant protein (rLOC387715).

Nano-flow liquid chromatography electrospray tandem mass spectrometry (nanoLC-ESI-MS/MS)

SDS-PAGE gel was stained by Coomassie Brilliant R-250 Blue. The gel band cut-out of bacterial expressed rLOC387715 was first destained with 50 mM Tris - 50% acetonitrile (pH 8.1) and reduced with 20 mM DTT in 50mM Tris (pH 8.1) at 55°C for 40 minutes, and then alkylated with 40 mM iodoacetamide at room temperature for 40 minutes in the dark. Proteins were digested in-situ with 30 μl enzyme solution (0.004 μg/μl porcine trypsin in 20 mM Tris - 0.0002% Zwittergent 3–16, pH 8.1) at 37°C overnight followed by peptide extraction with 60 μl of 2% trifluoroacetic acid, then 60 μl of acetonitrile. The pooled extracts were concentrated to less than 5 μl on a SpeedVac spinning concentrator (Savant Instruments, Holbrook NY) and then brought up in 0.1% formic acid for protein identification by nanoLC-ESI-MS/MS using a ThermoFinnigan LTQ Orbitrap Hybrid Mass Spectrometer (ThermoElectron Bremen, Germany) 12 coupled to an Eksigent nanoLC-2D HPLC system (Eksigent, Dublin, CA). The peptide mixture was loaded onto a 250-μl OPTI-PAK trap (Optimize Technologies, Oregon City, OR) custom packed with Michrom Magic C8 solid phase (Michrom Bioresources, Auburn, CA) and eluted with a 0.2 % formic acid-acetonitrile gradient through a Michrom packed tip capillary Magic C18 column (75 μm × 200 mm). The LTQ Orbitrap mass spectrometer experiment was set to perform a FT full scan from 380–1600 m/z with resolving power set at 60000 (400 m/z), followed by linear ion trap MS/MS scans on the top 3 ions. Dynamic exclusion was set to 2 and selected ions were placed on an exclusion list for 60 seconds. The MS/MS raw data were converted to DTA files using ThermoElectron Bioworks 3.2 and correlated to theoretical fragmentation patterns of tryptic peptide sequences from the NCBI nr database using Mascot™ (Matrix Sciences London, UK) search algorithm running on 10 node cluster 26. All searches were conducted with variable modifications allowing for carboxamidomethyl-cysteine, cysteic acid, proprionamide cysteine, methione sulphoxide, protein N-terminal acetylation, and deamidation of asparagine and glutamine. The search was restricted to trypsin-generated peptides allowing for 2 missed cleavages and was left open to all species. Peptide mass search tolerances are set to 10 ppm and fragment mass tolerance are set to ± 1.0 Daltons. Protein identifications were considered when Mascot search results gave at least two consensus peptides with individual peptide probability scores exceeding a threshold of 40, and ranking number one of all the hits for their respective MS/MS spectra.

Results

Synonymous modification of LOC387715 ORF cDNA sequence

The 2 NIH MGC cDNA clones (BC066349 and BC090924) of the “hypothetical” gene *LOC387715* all have a 107-aa encoding ORF, and the difference for the predicted protein sequence is that the 3rd aa for the latter is His rather than Arg that is present in the former (Fig. 1C and 1D). However, the aa sequence deduced from the ORF of BC066349 is exactly the same as the aa sequence (NCBI Accession #: XP_001131282) deduced from the predicted mRNA (NCBI Accession #: XM_001131282) that was obtained from the genomic DNA (Fig. 1B). Therefore, we choose to express the recombinant protein of *LOC387715* with same aa sequence as that of XP_001131282 for future studies.

A rare codon (RC) is a low-usage degenerate codon that is not only used rarely or infrequently but also decoded by rare (low-abundant) tRNA and/or other rare factors in an organism, which may quantitatively and/or qualitatively cause translational problems during a gene expression in the organism 4. Based on codon usage frequency, the abundance of the corresponding decoding tRNA, and the effect on quality and quantity of protein expression, *E. coli* RCs can be classified into 3 groups (I, IIa and IIb). There are 7, 6 and 7 Group-I, IIa and IIb *E. coli* RCs respectively, with the first 2 groups (Group-I and IIa) rather than the last group (IIb) of *E. coli* RCs having been widely reported to be TPC (translational problem causative) RCs 4. Moreover, rare codon clusters (RCC) consisting of RCs that are spaced by 0–5 common (non-TPC) codons, can exacerbate the RC-caused expression problems depending on the position of an RCC 2. Either BC066349 or XM_001131282 contains 5.6%, 15.0% and 12.2% Group-I, IIa and IIb 4 *E. coli* RCs, respectively (Table 1). In addition, in the ORF of BC066349 or XM_001131282, there are 3 and 4 simple *E. coli* RCCs where RCs are spaced by 1 or 2 non-TPC codons, and one complex *E. coli* RCC where RCs and RCCs are spaced by < 5 non-TPC codons (data not shown). These suggest that we would not obtain high and authentic expression of recombinant LOC387715 in *E. coli* if we directly employed the native ORF as a template.

In order to highly and authentically express rLOC387715 in *E. coli*, we directly used synonymous modification of the cDNA sequence present in the native ORF of BC066349 or XM_001131282 by chemical synthesis (Table 1, and Fig. 2A). In the modification, we synonymously substituted all 35 *E. coli* RCs (not only 6 Group-I and 16 Group-IIa RCs but also 13 Group-IIb RCs) with their corresponding optimal or sub-optimal common codons. In an mRNA sequence, a RBSCS (ribosome binding site complementary sequence, which may block translation initiation through mimicking the UCCU core sequence at the 3′ end of 16S rRNA and thus base-pairing with the Shine-Dalgarno or SD sequence required for initiation of translation) and a SDLS - a sequence similar to the SD sequence of a ribosome binding site (RBS) (which may bind to the UCCU core sequence at the 3′ end of 16S rRNA and block bacterial ribosomal binding to the correct SD site of an mRNA molecule for translation) may also cause no or undetectable expression of a foreign gene in *E. coli* 2. Therefore, when we replaced the ORF cDNA sequence, the optimal or sub-optimal degenerate common codon of an *E. coli* RC, e.g., substitution of TCC by TCT or AGC, was flexibly chosen in order to avoid forming of any new RBS and RBSCS sequence(s). Moreover, the synonymous modification automatically resulted in the elimination of the original 2 SDLS (aaagga and aggagcaaa) and 1 RBSCS (tcctt) present in the native ORF of *LOC387715*. Because codon adaptation index (CAI) value of a gene is a parameter often used to predict the protein expression level of the gene in the cells of an organism since it most often parallels to the levels of gene expression into protein products 37, we therefore calculated the CAI values before and after the modification. In *E. coli*, the CAI values for the native ORF of BC066349 or XM_001131282 and the synonymous modified *LOC387715* ORF cDNA sequence were 0.2647 and 0.8025 respectively, suggesting that the modified ORF sequence may be capable of highly expressing its encoded recombinant protein in *E. coli*.

High expression of recombinant protein rLOC387715 in E. coli

The chemically synthesized *LOC387715* ORF cDNA sequence – LOC387715(ORF)m, was first cloned in-frame into the expression vector pGS21a between the *Eco*RI and *Xho*I sites (Fig. 2B). The plasmid construct was then transformed in *E. coli* strain *BL21(DE3)*, and induction of the transformed bacteria with isopropylthio-β-D-galactoside (IPTG) at 1 mM at 37°C for 3h or overnight resulted in high expression of the recombinant protein rLOC387715 (His-tagged GST-LOC387715) – the amount was about 15% of the bacterial total proteins for 3h induction (Fig. 2C). Moreover, the plasmid construct was also transformed into *E. coli* strains *BL21* and *Rosetta 2(DE3)* (the latter carries a plasmid with the tRNA genes that decode seven major *E. coli* RCs - AGA, AGG, AUA, CUA, GGA, CCC, and CGG for improving the yield of full-length proteins 2), which obtained similar high levels of rLOC387715 expression (data not shown), suggesting that the expression of the rLOC387715 by LOC387715(ORF)m was not dependent on whether an expression host (*E. coli*) can express supplemental rare tRNAs.

The expression vector pGS21a encodes a His-tagged GST (glutathione-S-transferase) protein (Fig. 3, lanes of pGS21a, vector controls). Therefore, the bacterial expressed recombinant protein - rLOC387715 contained a His-tagged GST partner, and was able to be immunoreactive with anti-GST antibody (Fig. 3A, lane of pGS21a-LOC387715(ORF)m) and anti-6×His tag antibody (Fig. 3B, lane of pGS21a-LOC387715(ORF)m). The Western blotting results indirectly confirmed that rLOC387715 protein was highly expressed in *E. coli* with LOC387715(ORF)m.

Authentic expression of recombinant protein rLOC387715 in E. coli

Currently no anti-LOC387715 antibody is available for us to use Western blotting to confirm that we have authentically expressed the recombinant protein of *LOC387715*, that is to say, to determine if our bacterial expressed rLOC387715 contains the correct or the same aa sequence as that of XP_001131282. Therefore, we first digested the expressed rLOC387715 in-gel by trypsin, and then used sensitive nanoLC-ESI-MS/MS technology and Mascot search algorithm to identify the trypsin-derived peptides of rLOC387715. Mascot searches, conducted with fixed modification of carboxamidomethyl-cysteine (C) and variable modifications of methione sulphoxide (M) and protein N-terminal acetylation, found that (a) rLOC387715 contained a GST partner that was matched to the GST (EC 2.5.1.18) of fluke (*Schistosoma japonicum*) (NCBI Accession #: A26484) with a MOWSE score of 2133 (data not shown); and (b) rLOC387715 contained trypsin-peptides that were matched to the hypothetical protein LOC387715 (XP_001131282) with a MOWSE score of 1056, and 5 trypsin-derived peptides (peptides a, b, c, d and f, but not e) in Fig. 4, which covered 75% aa of LOC387715, were identified in rLOC387715. Furthermore, Mascot searches, conducted with variable modifications allowing for carboxamidomethyl-cysteine, cysteic acid, proprionamide cysteine, methione sulphoxide, protein N-terminal acetylation, and deamidation of asparagine and glutamine, were also performed, which found that all the 6 possible trypsin-digested peptides of LOC387715 (XP_001131282) (the MOWSE score for the match was 1023) were present in *E. coli* expressed rLOC387715 with the cysteine in peptide e (residues 71-88) was in the status of cysteic acid (Table 2 and Fig. 4). The 6 trypsin-derived peptides (a–f) covered 92% of the aa of LOC387715 (Fig. 4). The results suggest that rLOC387715 was authentically expressed in *E. coli*.

Discussion

The native ORF cDNA of *LOC387715* has a large number of *E. coli* RCs and RCCs, which may result in low and undetectable expression of the encoded protein in the bacterium because of (a) mRNA destabilization caused by impaired translation elongation at the RCs 47, (b) blocking or lowering of translation initiation 10,13,14,27, and (c) nascent protein degradation with 11,30,31 or without tm-RNA mediation 18. Specifically, an Arg RC – AGG (codon #107) preceding the stop codon TGA and a tandem RCC – AGGAGG (codons # 88–89) in the native *LOC387715* ORF cDNA implicate that it may cause low or undetectable expression of the encoded protein in *E. coli* by the mechanism of tm-RNA (SsrA)-mediated degradation of nascent proteins 11, lowering translation initiation through competing with Shine-Dalgarno (SD) sequence for the UCCU core sequence at the 3′-end of 16S rRNA on a ribosome 13,14 and other mechanisms such as low translation rate and ribosomal stalling 32,48. CAI analysis suggests that the native *LOC387715* ORF cDNA is a cDNA encoding low level of protein in *E. coli*. Moreover, *E. coli* RCs have been widely reported to cause the expression of a target gene in the bacterium not to produce the authentic target protein with the correct aa sequence, including misincorporation of wrong aa into the target protein 1,22,24,25,36,43, the synthesis of C-terminal truncated 6,23,28 or amino acid-deleted peptides or proteins 17, and the synthesis of frame-shifted target protein 7,9,20,39–41,44. Therefore, the native ORF cDNA of *LOC387715* was predicted not to highly and authentically express the recombinant protein of *LOC387715* in *E. coli*.

Our synonymously modified cDNA - LOC387715(ORF)m has the following characteristics: (a) it has no *E. coli* RCs and RCCs (all the 35 RCs including 6, 16 and 13 respective Group-I, IIa and IIb RCs, were synonymously substituted by their degenerate optimal or sub-optimal codons); (b) it has no *E. coli* SDLS and RBSCS sequences; and (c) it has a high calculated CAI value (0.8025). Therefore, theoretically it should highly express the encoded recombinant protein in *E. coli* without any aa changes. Practically, this is true because (a) SDS-PAGE and Western blotting demonstrated that plasmid pGS21a-LOC387715(ORF)m transformed *E. coli* highly expressed the recombinant protein - rLOC387715 (His-tagged GST-LOC387715), (b) Western blot with anti-GST antibody (Fig. 3A) did not find any detectable C-terminal truncated rLOC387715, and (c) nano-LC-ESI-MS/MS further confirmed that rLOC387715 was authentically expressed – the highly expressed rLOC387715 had the same aa sequence as that of XP_001131282. Therefore, the strategy of this study (Fig. 2A) for achieving high and authentic expression of rLOC387715 in *E. coli* by directly employing chemically synthesized synonymously modified cDNA - LOC387715(ORF)m rather than its native equivalent LOC387715(ORF), is a cost-effective one because (a) the price for gene/cDNA synthesis is minimal compared to other strategies, and (b) it avoids the classical clone-and-test method which is not only expensive and time-consuming, but does not use available knowledge to design an ORF that predicts high and authentic expression of the encoded recombinant protein.

The association of *LOC387715* and/or adjacent genes in chromosomal 10q26.13 region with AMD, and the strong interaction of *LOC387715* locus with smoking in AMD development 16,29,33,38, indicates that the “hypothetical” gene could play a critical role in AMD disease. A practical approach to determine if *LOC387715* is a real protein-encoding gene may be to use the antibody against the present authentic rLOC387715 protein to identify and/or isolate the possible native protein(s) encoded by this “gene” in human cells. Therefore, current expressed recombinant protein (a) may be directly used in the near future for the investigations related to the probing of the native protein(s) and function of *LOC387715*; and (b) may be further used as a valuable tool for studying or characterizing AMD.

Acknowledgments

This work was supported by EY014467, a departmental unrestricted award from Research to Prevent Blindness (RPB), the Foundation for Fighting Blindness (FFB), Presbyterian Hospital of Dallas/Presbyterian Healthcare Foundation, and the Eye Clinic & Department of Ophthalmology of Louisiana State University Health Sciences Center (LSUHSC). The authors thank Dr. Dennis O’Callaghan (Department of Microbiology and Immunology, LSUHSC, LA) for generously providing access to his core facility.

References

1. Calderone TL, Stevens RD, Oas TG. High-level misincorporation of lysine for arginine at AGA codons in a fusion protein expressed in Escherichia coli. J Mol Biol 1996;262:407–412. [PubMed: 8893852]
2. Chen D, Duggan C, Ganley JP, Kooragayala LM, Reden TB, Texada DE, Langford MP. Expression of enterovirus 70 capsid protein VP1 in Escherichia coli. Protein Expr Purif 2004;37:426–433. [PubMed: 15358366]
3. Chen D, Duggan C, Texada DE, Reden TB, Kooragayala LM, Langford MP. Immunogenicity of enterovirus 70 capsid protein VP1 and its non-overlapping N- and C-terminal fragments. Antiviral Res 2005;66:111–117. [PubMed: 15911028]
4. Chen D, Texada DE. Low-usage codons and rare codons of Escherichia coli. Gene Therapy and Molecular Biology 2006;10:1–12.
5. Chen D, Texada DE, Duggan C, Liang C, Reden TB, Kooragayala LM, Langford MP. Surface calreticulin mediates muramyl dipeptide-induced apoptosis in RK13 cells. J Biol Chem 2005;280:22425–22436. [PubMed: 15817475]
6. Choi AH, Basu M, McNeal MM, Bean JA, Clements JD, Ward RL. Intranasal administration of an Escherichia coli-expressed codon-optimized rotavirus VP6 protein induces protection in mice. Protein Expr Purif 2004;38:205–216. [PubMed: 15555936]
7. de Smit MH, van Duin J, van Knippenberg PH, van Eijk HG. CCC.UGA: a new site of ribosomal frameshifting in Escherichia coli. Gene 1994;143:43–47. [PubMed: 8200537]
8. Fisher SA, Abecasis GR, Yashar BM, Zareparsi S, Swaroop A, Iyengar SK, Klein BE, Klein R, Lee KE, Majewski J, Schultz DW, Klein ML, Seddon JM, Santangelo SL, Weeks DE, Conley YP, Mah TS, Schmidt S, Haines JL, Pericak-Vance MA, Gorin MB, Schulz HL, Pardi F, Lewis CM, Weber BH. Meta-analysis of genome scans of age-related macular degeneration. Hum Mol Genet 2005;14:2257–2264. [PubMed: 15987700]
9. Gurvich OL, Baranov PV, Gesteland RF, Atkins JF. Expression levels influence ribosomal frameshifting at the tandem rare arginine codons AGG_AGG and AGA_AGA in Escherichia coli. J Bacteriol 2005;187:4023–4032. [PubMed: 15937165]
10. Hall MN, Gabay J, Debarbouille M, Schwartz M. A role for mRNA secondary structure in the control of translation initiation. Nature 1982;295:616–618. [PubMed: 6799842]
11. Hayes CS, Bose B, Sauer RT. Stop codons preceded by rare arginine codons are efficient determinants of SsrA tagging in Escherichia coli. Proc Natl Acad Sci U S A 2002;99:3440–3445. [PubMed: 11891313]
12. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham CR. The Orbitrap: a new mass spectrometer. J Mass Spectrom 2005;40:430–443. [PubMed: 15838939]
13. Hu X, Shi Q, Yang T, Jackowski G. Specific replacement of consecutive AGG codons results in high-level expression of human cardiac troponin T in Escherichia coli. Protein Expr Purif 1996;7:289–293. [PubMed: 8860654]
14. Ivanov I, Alexandrova R, Dragulev B, Saraffova A, Abouhaidar MG. Effect of tandemly repeated AGG triplets on the translation of CAT-mRNA in E. coli. FEBS Lett 1992;307:173–176. [PubMed: 1379538]
15. Iyengar SK, Song D, Klein BE, Klein R, Schick JH, Humphrey J, Millard C, Liptak R, Russo K, Jun G, Lee KE, Fijal B, Elston RC. Dissection of genomewide-scan data in extended families reveals a major locus and oligogenic susceptibility for age-related macular degeneration. Am J Hum Genet 2004;74:20–39. [PubMed: 14691731]
16. Jakobsdottir J, Conley YP, Weeks DE, Mah TS, Ferrell RE, Gorin MB. Susceptibility genes for age-related maculopathy on chromosome 10q26. Am J Hum Genet 2005;77:389–407. [PubMed: 16080115]
17. Kane JF, Violand BN, Curran DF, Staten NR, Duffin KL, Bogosian G. Novel in-frame two codon translational hop during synthesis of bovine placental lactogen in a recombinant strain of Escherichia coli. Nucleic Acids Res 1992;20:6707–6712. [PubMed: 1480491]
18. Kapust RB, Routzahn KM, Waugh DS. Processive degradation of nascent polypeptides, triggered by tandem AGA codons, limits the accumulation of recombinant tobacco etch virus protease in Escherichia coli BL21(DE3). Protein Expr Purif 2002;24:61–70. [PubMed: 11812224]

19. Kenealy SJ, Schmidt S, Agarwal A, Postel EA, De La Paz MA, Pericak-Vance MA, Haines JL. Linkage analysis for age-related macular degeneration supports a gene on chromosome 10q26. Mol Vis 2004;10:57–61. [PubMed: 14758336]

20. Li Z, Stahl G, Farabaugh PJ. Programmed +1 frameshifting stimulated by complementarity between a downstream mRNA sequence and an error-correcting region of rRNA. RNA 2001;7:275–284. [PubMed: 11233984]

21. Majewski J, Schultz DW, Weleber RG, Schain MB, Edwards AO, Matise TC, Acott TS, Ott J, Klein ML. Age-related macular degeneration--a genome scan in extended families. Am J Hum Genet 2003;73:540–550. [PubMed: 12900797]

22. McNulty DE, Claffee BA, Huddleston MJ, Kane JF. Mistranslational errors associated with the rare arginine codon CGG in Escherichia coli. Protein Expr Purif 2003;27:365–374. [PubMed: 12597898]

23. Misra R, Reeves P. Intermediates in the synthesis of TolC protein include an incomplete peptide stalled at a rare Arg codon. Eur J Biochem 1985;152:151–155. [PubMed: 3899641]

24. Parker J. Errors and alternatives in reading the universal genetic code. Microbiol Rev 1989;53:273–298. [PubMed: 2677635]

25. Parker J, Friesen JD. “Two out of three” codon reading leading to mistranslation in vivo. Mol Gen Genet 1980;177:439–445. [PubMed: 6768967]

26. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999;20:3551–3567. [PubMed: 10612281]

27. Pohlner J, Meyer TF, Jalajakumari MB, Manning PA. Nucleotide sequence of ompV, the gene for a major Vibrio cholerae outer membrane protein. Mol Gen Genet 1986;205:494–500. [PubMed: 3031428]

28. Ramachandiran V, Kramer G, Horowitz PM, Hardesty B. Single synonymous codon substitution eliminates pausing during chloramphenicol acetyl transferase synthesis on Escherichia coli ribosomes in vitro. FEBS Lett 2002;512:209–212. [PubMed: 11852081]

29. Rivera A, Fisher SA, Fritsche LG, Keilhauer CN, Lichtner P, Meitinger T, Weber BH. Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. Hum Mol Genet 2005;14:3227–3236. [PubMed: 16174643]

30. Roche ED, Sauer RT. SsrA-mediated peptide tagging caused by rare codons and tRNA scarcity. EMBO J 1999;18:4579–4589. [PubMed: 10449423]

31. Roche ED, Sauer RT. Identification of endogenous SsrA-tagged proteins reveals tagging at positions corresponding to stop codons. J Biol Chem 2001;276:28509–28515. [PubMed: 11373298]

32. Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G. Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. J Bacteriol 1993;175:716–722. [PubMed: 7678594]

33. Schmidt S, Hauser MA, Scott WK, Postel EA, Agarwal A, Gallins P, Wong F, Chen YS, Spencer K, Schnetz-Boutaud N, Haines JL, Pericak-Vance MA. Cigarette smoking strongly modifies the association of LOC387715 and age-related macular degeneration. Am J Hum Genet 2006;78:852–864. [PubMed: 16642439]

34. Schmidt S, Scott WK, Postel EA, Agarwal A, Hauser ER, De La Paz MA, Gilbert JR, Weeks DE, Gorin MB, Haines JL, Pericak-Vance MA. Ordered subset linkage analysis supports a susceptibility locus for age-related macular degeneration on chromosome 16p12. BMC Genet 2004;5:18. [PubMed: 15238159]

35. Seddon JM, Santangelo SL, Book K, Chong S, Cote J. A genomewide scan for age-related macular degeneration provides evidence for linkage to several chromosomal regions. Am J Hum Genet 2003;73:780–790. [PubMed: 12945014]

36. Seetharam R, Heeren RA, Wong EY, Braford SR, Klein BK, Aykent S, Kotts CE, Mathis KJ, Bishop BF, Jennings MJ. Mistranslation in IGF-1 during over-expression of the protein in Escherichia coli using a synthetic gene containing low frequency codons. Biochem Biophys Res Commun 1988;155:518–523. [PubMed: 3137938]

37. Sharp PM, Li WH. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987;15:1281–1295. [PubMed: 3547335]

38. Shastry BS. Further support for the common variants in complement factor H (Y402H) and LOC387715 (A69S) genes as major risk factors for the exudative age-related macular degeneration. Ophthalmologica 2006;220:291–295. [PubMed: 16954704]

39. Shu P, Dai H, Mandecki W, Goldman E. CCC CGA is a weak translational recoding site in Escherichia coli. Gene 2004;343:127–132. [PubMed: 15563838]

40. Spanjaard RA, Chen K, Walker JR, van Duin J. Frameshift suppression at tandem AGA and AGG codons by cloned tRNA genes: assigning a codon to argU tRNA and T4 tRNA(Arg). Nucleic Acids Res 1990;18:5031–5036. [PubMed: 2205835]

41. Spanjaard RA, van Duin J. Translation of the sequence AGG-AGG yields 50% ribosomal frameshift. Proc Natl Acad Sci U S A 1988;85:7967–7971. [PubMed: 3186700]

42. Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Buetow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Brownstein MJ, Usdin TB, Toshiyuki S, Carninci P, Prange C, Raha SS, Loquellano NA, Peters GJ, Abramson RD, Mullahy SJ, Bosak SA, McEwan PJ, McKernan KJ, Malek JA, Gunaratne PH, Richards S, Worley KC, Hale S, Garcia AM, Gay LJ, Hulyk SW, Villalon DK, Muzny DM, Sodergren EJ, Lu X, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madan A, Young AC, Shevchenko Y, Bouffard GG, Blakesley RW, Touchman JW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Krzywinski MI, Skalska U, Smailus DE, Schnerch A, Schein JE, Jones SJ, Marra MA. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci U S A 2002;99:16899–16903. [PubMed: 12477932]

43. Tsai F, Curran JF. tRNA(2Gln) mutants that translate the CGA arginine codon as glutamine in Escherichia coli. RNA 1998;4:1514–1522. [PubMed: 9848650]

44. Vilbois F, Caspers P, da Prada M, Lang G, Karrer C, Lahm HW, Cesura AM. Mass spectrometric analysis of human soluble catechol O-methyltransferase expressed in Escherichia coli. Identification of a product of ribosomal frameshifting and of reactive cysteines involved in S-adenosyl-L-methionine binding. Eur J Biochem 1994;222:377–386. [PubMed: 8020475]

45. Weeks DE, Conley YP, Tsai HJ, Mah TS, Rosenfeld PJ, Paul TO, Eller AW, Morse LS, Dailey JP, Ferrell RE, Gorin MB. Age-related maculopathy: an expanded genome-wide scan with evidence of susceptibility loci within the 1q31 and 17q25 regions. Am J Ophthalmol 2001;132:682–692. [PubMed: 11704029]

46. Weeks DE, Conley YP, Tsai HJ, Mah TS, Schmidt S, Postel EA, Agarwal A, Haines JL, Pericak-Vance MA, Rosenfeld PJ, Paul TO, Eller AW, Morse LS, Dailey JP, Ferrell RE, Gorin MB. Age-related maculopathy: a genomewide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions. Am J Hum Genet 2004;75:174–189. [PubMed: 15168325]

47. Wu X, Jornvall H, Berndt KD, Oppermann U. Codon optimization reveals critical factors for high level expression of two rare codon genes in Escherichia coli: RNA stability and secondary structure but not tRNA abundance. Biochem Biophys Res Commun 2004;313:89–96. [PubMed: 14672702]

48. Zahn K, Landy A. Modulation of lambda integrase synthesis by rare arginine tRNA. Mol Microbiol 1996;21:69–76. [PubMed: 8843435]

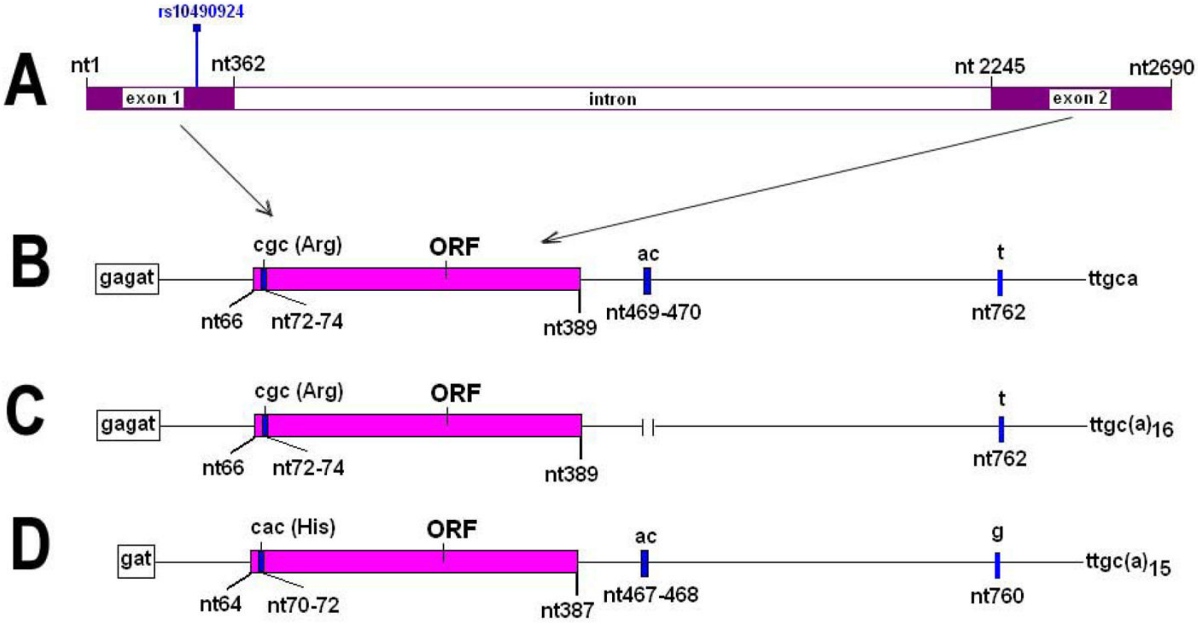

Figure 1. The structures of *LOC387715* and its predicted mRNA and 2 cDNA clones

A. Hypothetical *LOC387715*. It contains one intron and 2 exons (nt1-362, and nt2245–2690) at both ends. **B.** The predicted mRNA of *LOC387715* (XM_373477). **C.** The NIH MGC cDNA clone BC066349 of *LOC387715*. **D.** The NIH MGC cDNA clone BC090924 of *LOC387715*. The major difference between XM_373477 and BC066349 is that the later had a deletion of 2 adjacent nucleotides (ac) at nt469–470. The major difference between XM_373477 and BC090924 is that the later at nt71 and nt760 has different nucleotides (a rather than g for nt71 - which causes the 3rd aa of the predicted protein to be His rather than Arg, and g rather than t for nt760). ORF – open reading frame, nt – nucleotides, and rs10490924 – the coding polymorphism that is associated with the risk of AMD.

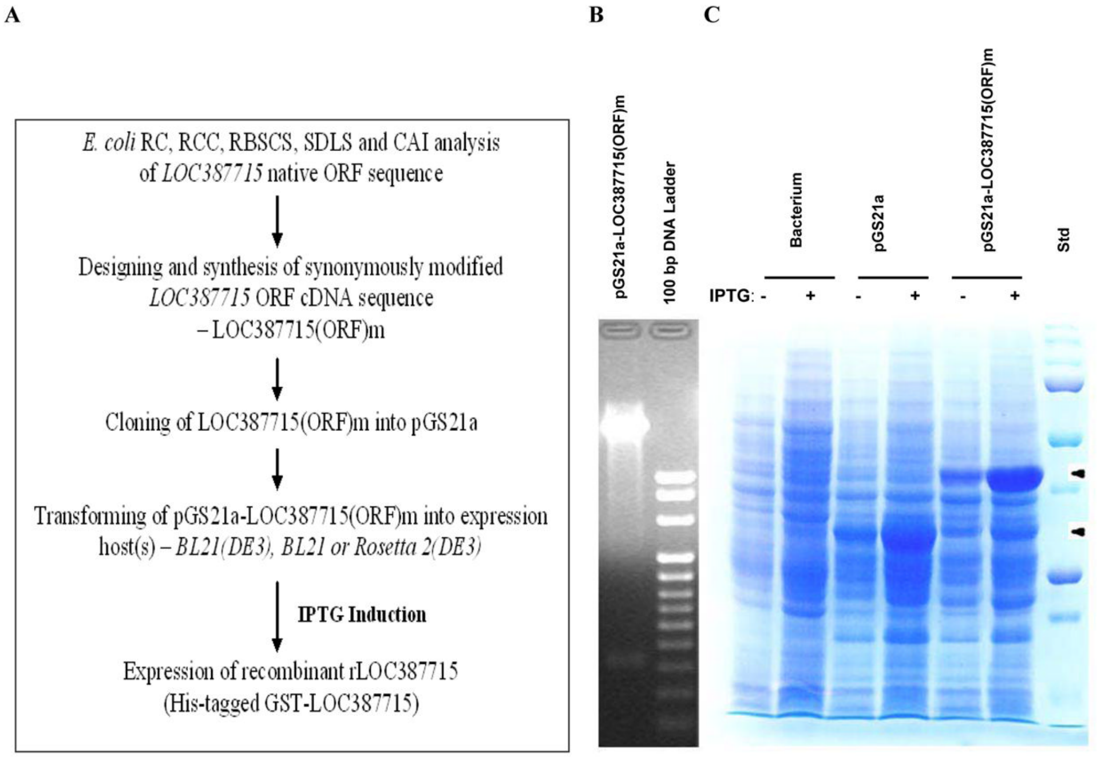
A
B
C
E. coli RC, RCC, RBSCS, SDLS and CAI analysis of LOC387715 native ORF sequence
Designing and synthesis of synonymously modified LOC387715 ORF cDNA sequence – LOC387715(ORF)m
Cloning of LOC387715(ORF)m into pGS21a
Transforming of pGS21a-LOC387715(ORF)m into expression host(s) – BL21(DE3), BL21 or Rosetta 2(DE3)
IPTG Induction
Expression of recombinant rLOC387715 (His-tagged GST-LOC387715)
pGS21a-LOC387715(ORF)m
100 bp DNA Ladder
Bacterium
pGS21a
pGS21a-LOC387715(ORF)m
Std
IPTG: - + - + - +

Figure 2. The expression of recombinant *LOC387715* protein (rLOC387715)

A. Schematic diagram of the strategy and process for high and authentical expression of rLOC387715 in *E. coli*. **B.** *EcoR I* + *Xho I* digestion of pGS21a-LOC387715(ORF)m, showing the expected size of insert (336 bp). **C.** SDS-PAGE, showing the high expression of the recombinant protein in bacterium *BL21(DE3)* strain after induction with IPTG (1mM at 37°C for 3 h). The recombinant protein rLOC387715 was about 15% of the bacterial total proteins. Top arrow, the recombinant protein – His-tagged GST-LOC387715 fusion protein that was expressed from pGS21a-LOC387715(ORF)m; bottom arrow, His-tagged GST that was expressed from the vector pGS21a. Std – protein standards.

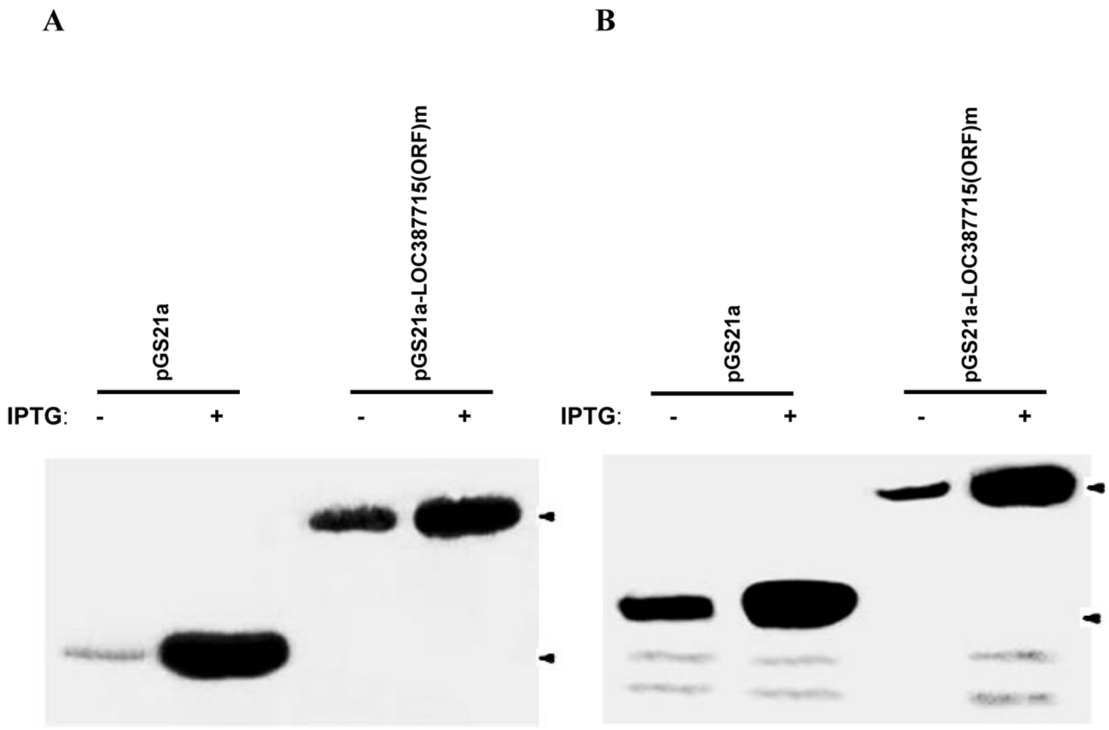

Figure 3. Western blotting of *E. coli* expressed recombinant *LOC387715* protein (rLOC387715)

A. Detection by anti-GST antibody. **B.** Detection by anti-6xHis antibody. High expression of the recombinant protein was obtained in bacterium *BL21(DE3)* strain by the induction of IPTG (1 mM at 37°C for 3 h). Top arrow, the recombinant protein – His-tagged GST-LOC387715 fusion protein that was expressed from pGS21a-LOC387715(ORF)m; bottom arrow, His-tagged GST that was expressed from the vector pGS21a.

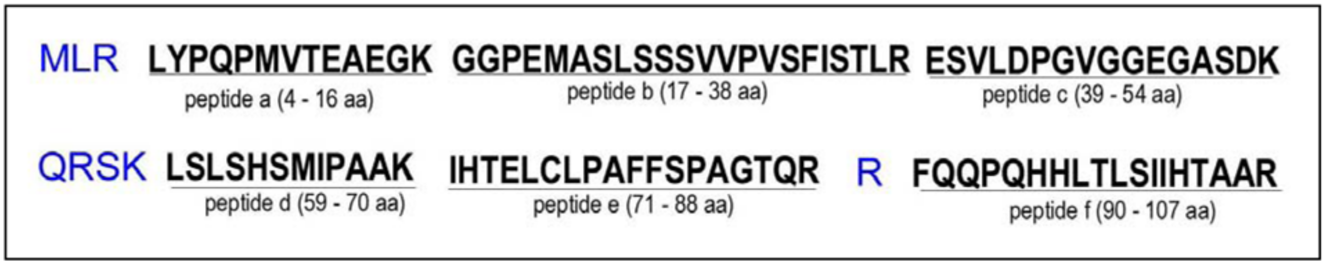

Figure 4. The assembly of mass spectrometry identified trypsin digestion-produced peptides of *E. coli* expressed recombinant protein rLOC387715

Amino acid R, and small peptides MLR, QR and SK that were produced by porcine trypsin digestion were not identified by nanoLC-ESI-MS/MS and highlighted in blue. Only LOC387715 protein part was shown in the figure.

Table 1

Synonymous modification of the *E. coli* rare codons present in the native ORF of *LOC387715* cDNA clones.

Codon	RC-group[a]	Substituted by	Amino acid encoded	Total #	Position[b]
AGG	I	CGT	Arg	4	56, 88, 89, 107
CGA	I	CGT	Arg	1	38
CTA	I	CTG	Leu	1	4
ACA	IIa	ACC	Thr	1	98
CCT	IIa	CCG	Pro	5	19, 30, 44, 83, 93
TCA	IIa	TCT	Ser	1	62
GGA	IIa	GGT	Gly	6	7, 17, 45, 48, 50, 85
AGT	IIa	TCT	Ser	2	23, 52
TCG	IIa	AGC	Ser	1	27
CCA	IIb	CCG	Pro	3	6, 67, 78
TCC	IIb	TCT	Ser	2	26, 64
TCC	IIb	AGC	Ser	3	25, 32, 35
GGG	IIb	GGC	Gly	2	15, 18
CTC	IIb	CTG	Leu	1	75
TTA	IIb	CTG	Leu	2	61, 77

[a] RC-group is rare codon group according to Chen and Texada 4.

[b] Position is expressed as the codon numbers in the cDNA clones BC066349 and BC090924 of *LOC387715*.

Table 2

Identification of trypsin digestion-derived peptides of rLOC387715 by nanoLC-ESI-MS/MS.

Peptide #	Residue# Start-end	Peptide ion Observed m/z	Molecular Weight (Mr)		Delta	Peptide sequence/Modification (aa)	Score[*]
			Expected	Calculated			
a	4 – 16	696.3477	1390.68	1390.68	0.0006	R.LYPGPMVTEAEGK.G	62
a	4 – 16	704.3453	1406.68	1406.675	0.0010	R.LYPGPMVTEAEGK.G/Oxidation (M)	55
b	17 – 38	1111.0804	2220.15	2220.146	0.0004	K.GGPEMASLSSSVVPVSFISTLR.E	112
b	17 – 38	1119.0799	2236.14	2236.141	0.0004	K.GGPEMASLSSSVVPVSFISTLR.E/Oxidation (M)	90
c	39 – 54	758.8605	1515.71	1515.705	0.0012	R.ESVLDPGVGGEGASDK.Q	87
d	59 – 70	627.8472	1253.68	1253.68	−0.0004	K.LSLSHSMIPAAK.I	64
d	59 – 70	635.8446	1269.67	1269.675	−0.0004	K.LSLSHSMIPAAK.I/Oxidation (M)	64
e	71 – 88	1018.4979	2034.98	2034.983	−0.0020	K.IHTELCLPAFFSPAGTQR.R/Cysteic-acid (C)	74
f	90 – 107	700.0483	2097.12	2097.123	0.0000	R.FQQPQHHLTLSIIHTAAR	52

[*] The score is peptide probability score.