Tin Sinh học
Bioinformatics

Bài thực hành 1. Hướng dẫn sử dụng NCBI

Tài liệu tham khảo

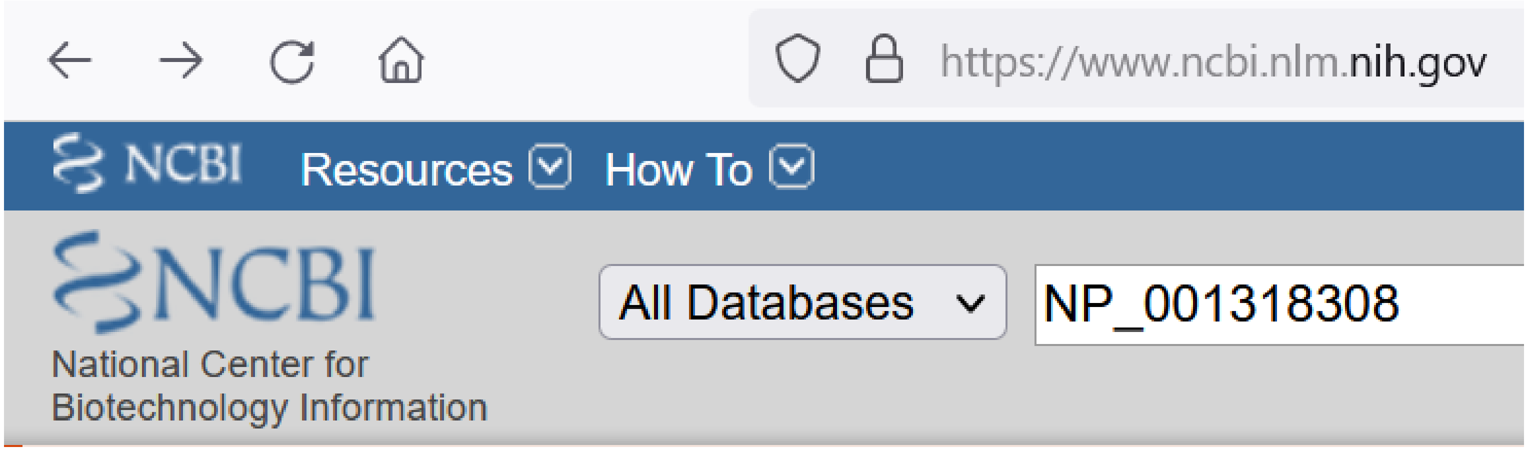
Nicholas James Provart, Bioinformatic Methods I, Coursera, University of Toronto, 2021.

Nội dung báo cáo

- Mô tả các bước thực hành.
- Trả lời các câu hỏi.
- Chụp màn hình kết quả thu được.
- Download các file trình tự.
- Nộp các kết quả trên vào Assignment trên Microsoft Team.

Bước 1. Lấy 1 trình tự DNA

- On the Search NCBI Portal page, search “All Databases” for your given protein sequence again using the Accession number.
- Using the protein from the first part of this lab, we would search for ***NP_001318308***.

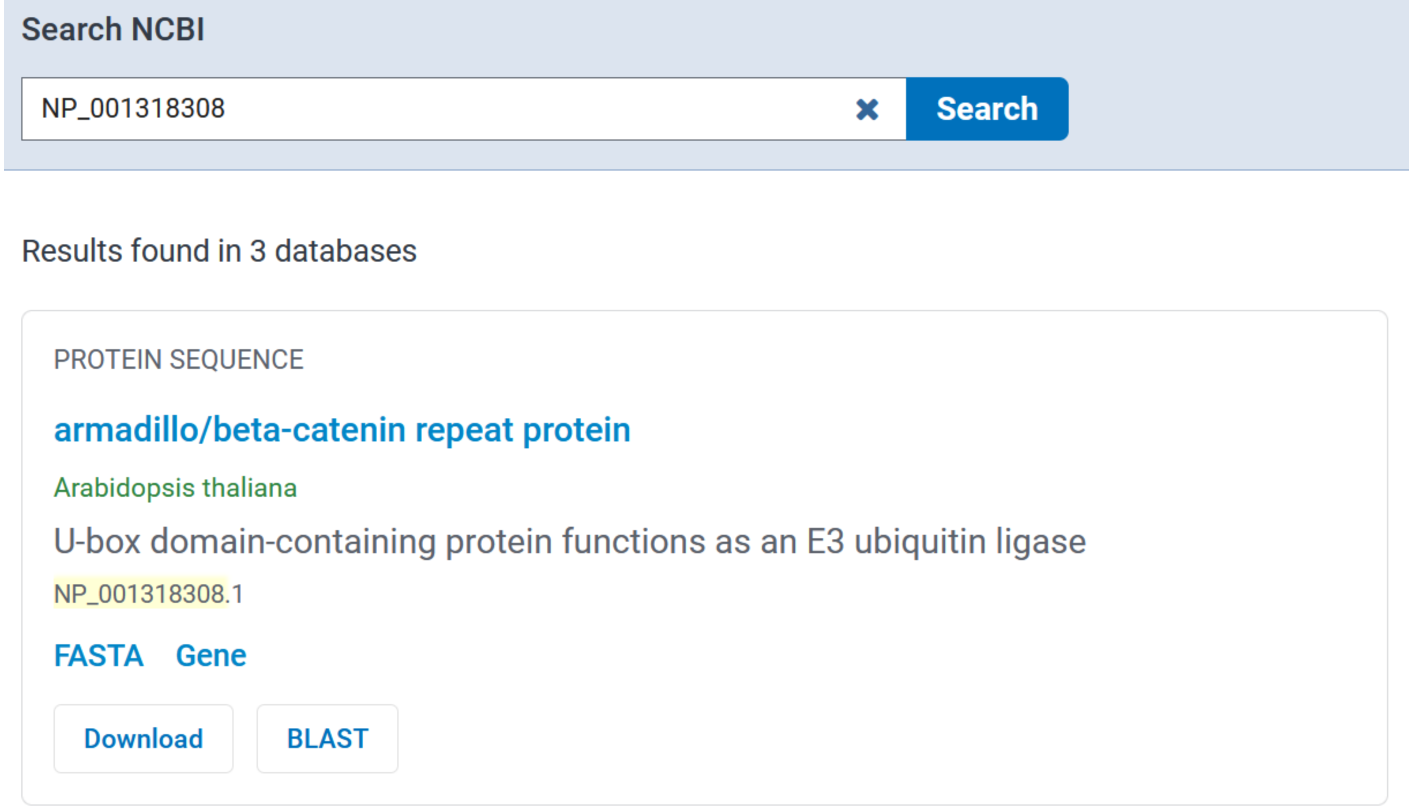


https://www.ncbi.nlm.nih.gov

All Databases

NP_001318308

- The first page that comes up is the summary page. Once you’re on this page you can move to the database of interest.
- In this case you probably don't have hits in too many databases since you had a very specific search.



Search NCBI

NP_001318308

Results found in 3 databases

PROTEIN SEQUENCE

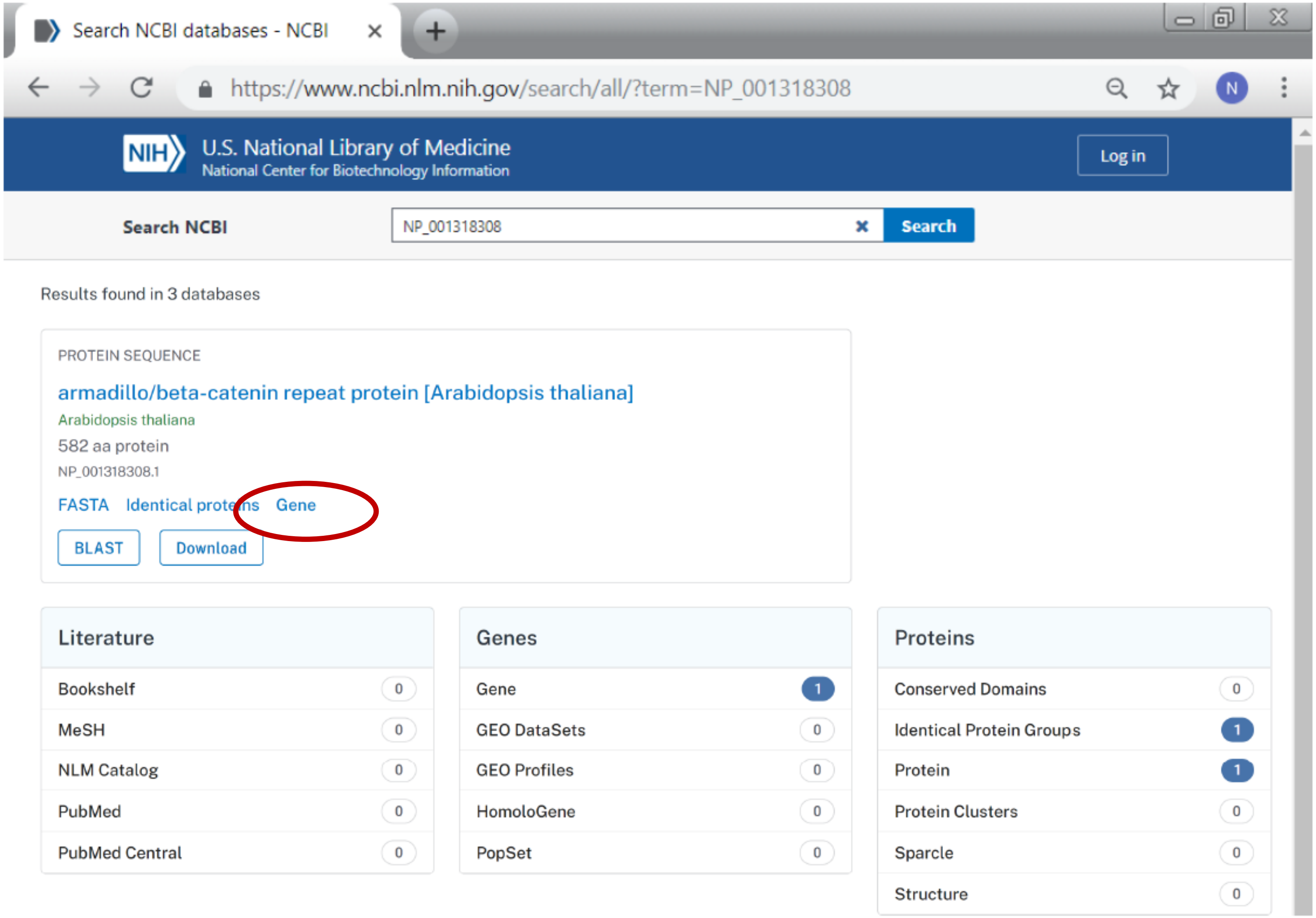
armadillo/beta-catenin repeat protein

Arabidopsis thaliana

U-box domain-containing protein functions as an E3 ubiquitin ligase

NP_001318308.1

Search NCBI portal queried for NP_001318308 with Gene results highlighted



Search NCBI databases - NCBI
https://www.ncbi.nlm.nih.gov/search/all/?term=NP_001318308
NIH U.S. National Library of Medicine
National Center for Biotechnology Information

Search NCBI: NP_001318308

Results found in 3 databases

PROTEIN SEQUENCE

armadillo/beta-catenin repeat protein [Arabidopsis thaliana]

Arabidopsis thaliana

582 aa protein

NP_001318308.1

FASTA
Identical proteins
Gene
BLAST
Download

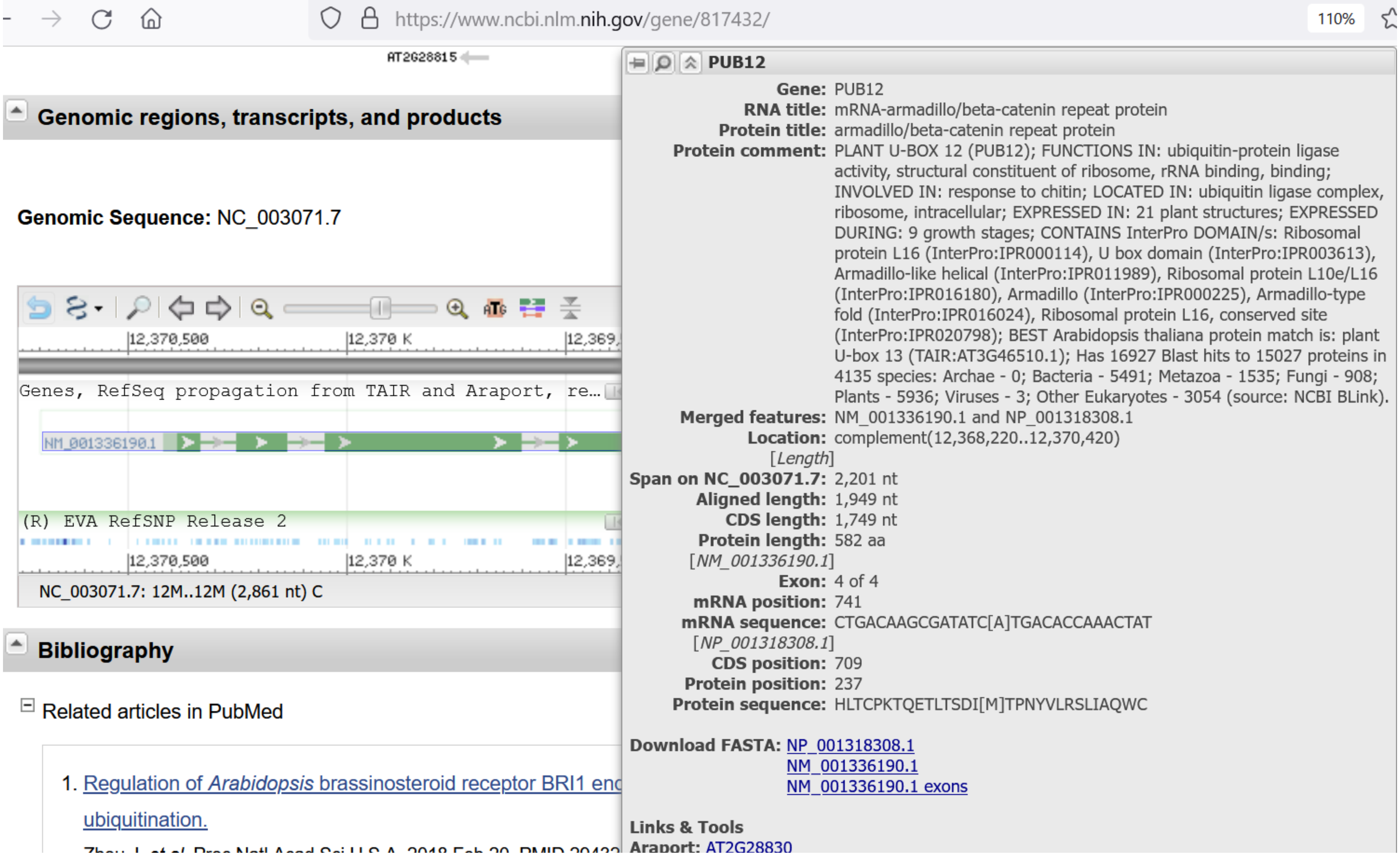
Literature	
Bookshelf	0
MeSH	0
NLM Catalog	0
PubMed	0
PubMed Central	0

Genes	
Gene	1
GEO DataSets	0
GEO Profiles	0
HomoloGene	0
PopSet	0

Proteins	
Conserved Domains	0
Identical Protein Groups	1
Protein	1
Protein Clusters	0
Sparcle	0
Structure	0

https://www.ncbi.nlm.nih.gov/search/all/?term=NP_001318308

- Try clicking the Gene link. Does the Gene page give you the gene sequence alone?
- What do you get instead?
- Note the context specific link menus that pop up when you hover over the graphic of the gene with your mouse pointer.
- You can click on the green boxes denoting the exons of the gene to get links to various sequences and analyses associated with the gene.
- Note that the green track is a composite of the mRNA and CDS tracks – click on either the NM_ (mRNA) or NP_(protein) number to see the deconvolution of the green track



Genomic regions, transcripts, and products

Genomic Sequence: NC_003071.7

NC_003071.7: 12M..12M (2,861 nt) C

Bibliography

⊟ Related articles in PubMed

1. Regulation of *Arabidopsis* brassinosteroid receptor BRI1 end ubiquitination.

PUB12

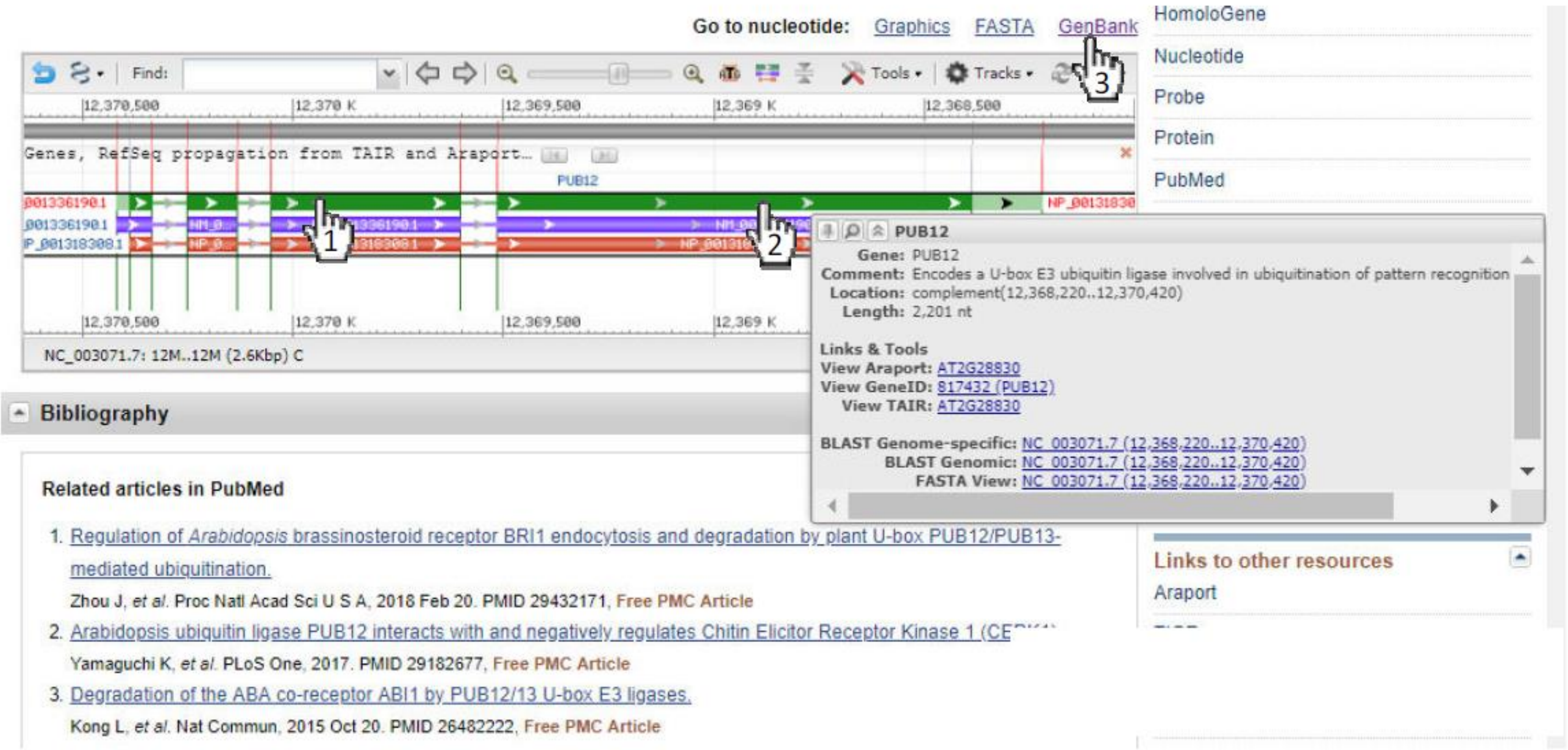
Gene: PUB12
RNA title: mRNA-armadillo/beta-catenin repeat protein
Protein title: armadillo/beta-catenin repeat protein
Protein comment: PLANT U-BOX 12 (PUB12); FUNCTIONS IN: ubiquitin-protein ligase activity, structural constituent of ribosome, rRNA binding, binding; INVOLVED IN: response to chitin; LOCATED IN: ubiquitin ligase complex, ribosome, intracellular; EXPRESSED IN: 21 plant structures; EXPRESSED DURING: 9 growth stages; CONTAINS InterPro DOMAIN/s: Ribosomal protein L16 (InterPro:IPR000114), U box domain (InterPro:IPR003613), Armadillo-like helical (InterPro:IPR011989), Ribosomal protein L10e/L16 (InterPro:IPR016180), Armadillo (InterPro:IPR000225), Armadillo-type fold (InterPro:IPR016024), Ribosomal protein L16, conserved site (InterPro:IPR020798); BEST Arabidopsis thaliana protein match is: plant U-box 13 (TAIR:AT3G46510.1); Has 16927 Blast hits to 15027 proteins in 4135 species: Archae - 0; Bacteria - 5491; Metazoa - 1535; Fungi - 908; Plants - 5936; Viruses - 3; Other Eukaryotes - 3054 (source: NCBI BLink).
Merged features: NM_001336190.1 and NP_001318308.1
Location: complement(12,368,220..12,370,420)
[*Length*]
Span on NC_003071.7: 2,201 nt
Aligned length: 1,949 nt
CDS length: 1,749 nt
Protein length: 582 aa
[*NM_001336190.1*]
Exon: 4 of 4
mRNA position: 741
mRNA sequence: CTGACAAGCGATATC[A]TGACACCAAACTAT
[*NP_001318308.1*]
CDS position: 709
Protein position: 237
Protein sequence: HLTCPKTQETLTSDI[M]TPNYVLRSLIAQWC

Download FASTA: NP_001318308.1
NM_001336190.1
NM_001336190.1 exons

Links & Tools
Araport: AT2G28830

https://www.ncbi.nlm.nih.gov/gene/817432/

Part of the Gene page for NP_001318308



Go to nucleotide: Graphics FASTA GenBank

HomoloGene

Nucleotide

Probe

Protein

PubMed

PUB12

Gene: PUB12
Comment: Encodes a U-box E3 ubiquitin ligase involved in ubiquitination of pattern recognition
Location: complement(12,368,220..12,370,420)
Length: 2,201 nt

Links & Tools
View Araport: AT2G28830
View GeneID: 817432 (PUB12)
View TAIR: AT2G28830

BLAST Genome-specific: NC_003071.7 (12,368,220..12,370,420)
BLAST Genomic: NC_003071.7 (12,368,220..12,370,420)
FASTA View: NC_003071.7 (12,368,220..12,370,420)

Bibliography

Related articles in PubMed

1. Regulation of *Arabidopsis* brassinosteroid receptor BRI1 endocytosis and degradation by plant U-box PUB12/PUB13-mediated ubiquitination.
   Zhou J, *et al.* Proc Natl Acad Sci U S A. 2018 Feb 20. PMID 29432171, Free PMC Article
2. Arabidopsis ubiquitin ligase PUB12 interacts with and negatively regulates Chitin Elicitor Receptor Kinase 1 (CE
   Yamaguchi K, *et al.* PLoS One. 2017. PMID 29182677, Free PMC Article
3. Degradation of the ABA co-receptor ABI1 by PUB12/13 U-box E3 ligases.
   Kong L, *et al.* Nat Commun. 2015 Oct 20. PMID 26482222, Free PMC Article

Links to other resources

Araport

https://www.ncbi.nlm.nih.gov/gene/817432/

Part of the Gene page for NP_001318308

- Showing pop-up to sequence links.

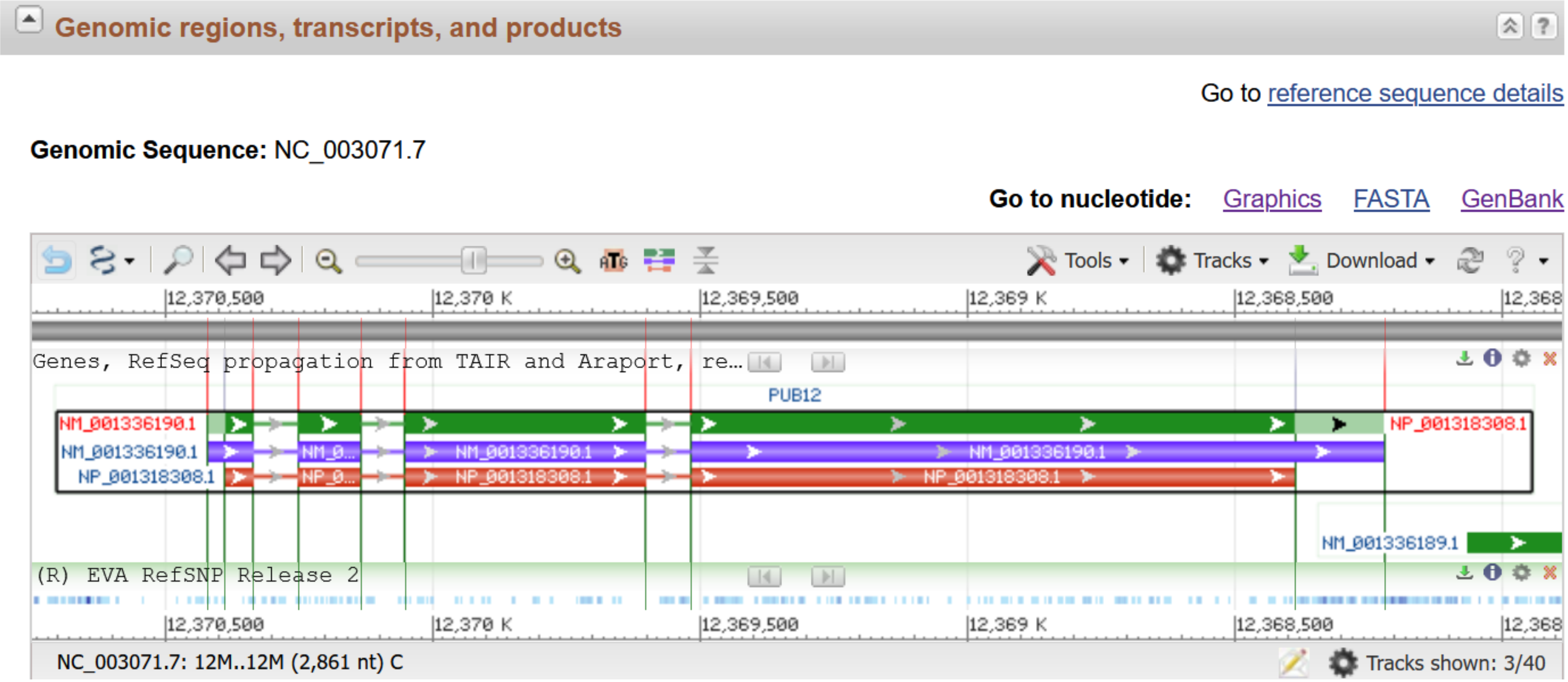
1. Click the green bars to make mRNA and protein tracks appear;

2. Hover over the mRNA track to see info panel;

3. Click “Genbank” link to see Genbank record for the genomic region for this gene.

https://www.ncbi.nlm.nih.gov/nuccore/NC_003071.7?report=genbank&from=12368220&to=12370420&strand=true

1. Click the green bars to make mRNA and protein tracks appear



Genomic regions, transcripts, and products

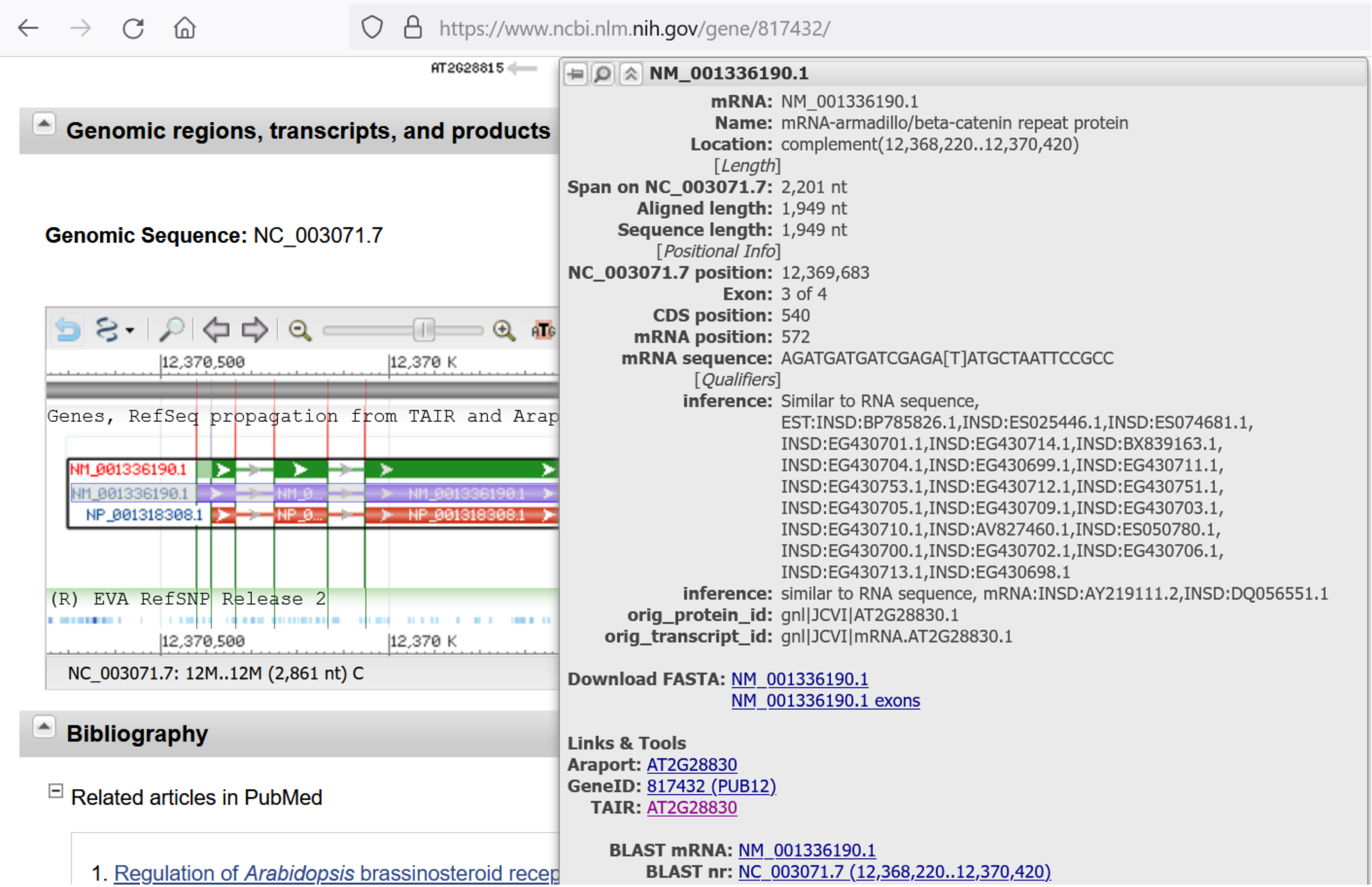
Go to reference sequence details

Genomic Sequence: NC_003071.7

Go to nucleotide: Graphics FASTA GenBank

https://www.ncbi.nlm.nih.gov/gene/817432/

2. Hover over the mRNA track to see info panel



https://www.ncbi.nlm.nih.gov/gene/817432/

AT2G28815

Genomic regions, transcripts, and products

Genomic Sequence: NC_003071.7

NC_003071.7: 12M..12M (2,861 nt) C

Bibliography

Related articles in PubMed

1. Regulation of *Arabidopsis* brassinosteroid recep

NM_001336190.1

mRNA: NM_001336190.1
Name: mRNA-armadillo/beta-catenin repeat protein
Location: complement(12,368,220..12,370,420)
[*Length*]
Span on NC_003071.7: 2,201 nt
Aligned length: 1,949 nt
Sequence length: 1,949 nt
[*Positional Info*]
NC_003071.7 position: 12,369,683
Exon: 3 of 4
CDS position: 540
mRNA position: 572
mRNA sequence: AGATGATGATCGAGA[T]ATGCTAATTCCGCC
[*Qualifiers*]
inference: Similar to RNA sequence, EST:INSD:BP785826.1,INSD:ES025446.1,INSD:ES074681.1, INSD:EG430701.1,INSD:EG430714.1,INSD:BX839163.1, INSD:EG430704.1,INSD:EG430699.1,INSD:EG430711.1, INSD:EG430753.1,INSD:EG430712.1,INSD:EG430751.1, INSD:EG430705.1,INSD:EG430709.1,INSD:EG430703.1, INSD:EG430710.1,INSD:AV827460.1,INSD:ES050780.1, INSD:EG430700.1,INSD:EG430702.1,INSD:EG430706.1, INSD:EG430713.1,INSD:EG430698.1
inference: similar to RNA sequence, mRNA:INSD:AY219111.2,INSD:DQ056551.1
orig_protein_id: gnl|JCVI|AT2G28830.1
orig_transcript_id: gnl|JCVI|mRNA.AT2G28830.1

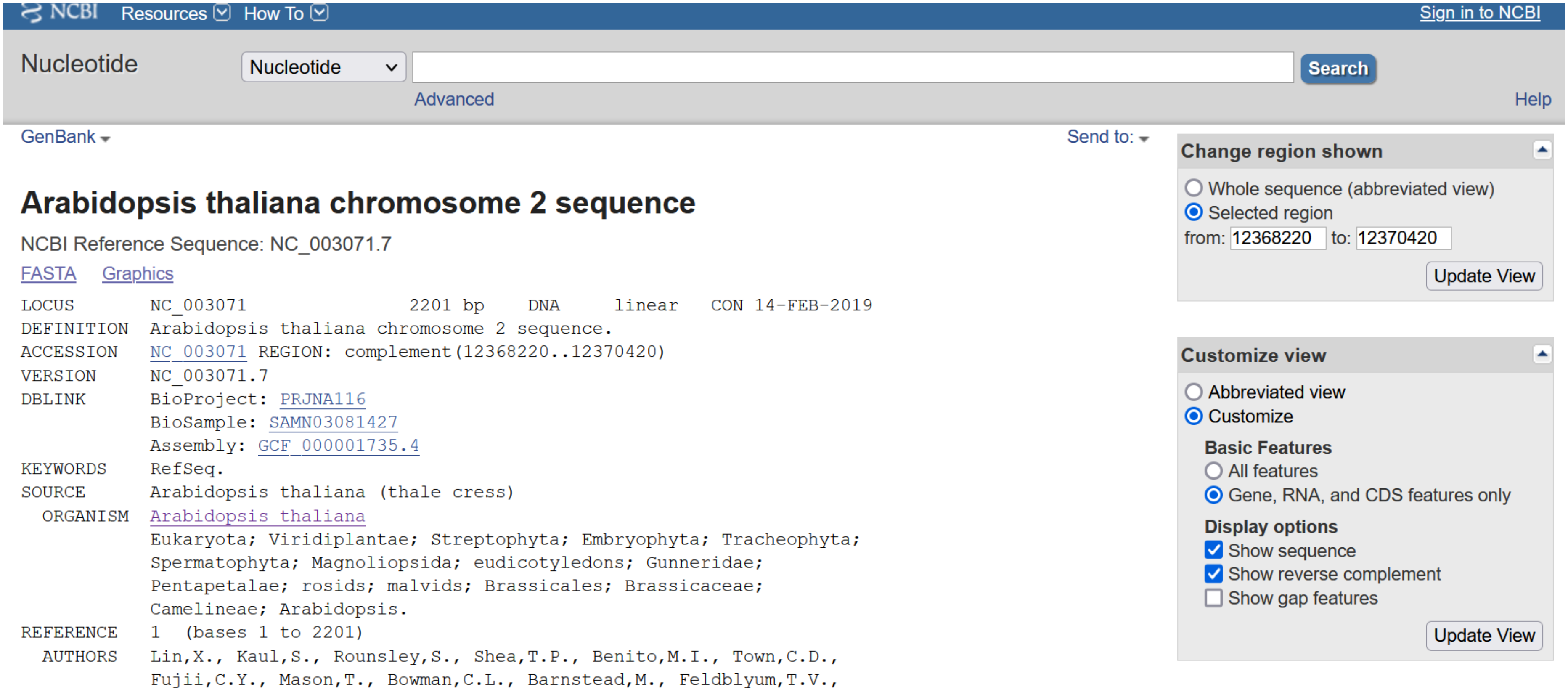
Download FASTA: NM_001336190.1
NM_001336190.1 exons

Links & Tools
Araport: AT2G28830
GeneID: 817432 (PUB12)
TAIR: AT2G28830

BLAST mRNA: NM_001336190.1
BLAST nr: NC_003071.7 (12,368,220..12,370,420)

https://www.ncbi.nlm.nih.gov/gene/817432/

3. Click “Genbank” link to see Genbank record for the genomic region for this gene



Arabidopsis thaliana chromosome 2 sequence

NCBI Reference Sequence: NC_003071.7

```
LOCUS       NC_003071               2201 bp    DNA     linear   CON 14-FEB-2019
DEFINITION  Arabidopsis thaliana chromosome 2 sequence.
ACCESSION   NC_003071 REGION: complement(12368220..12370420)
VERSION     NC_003071.7
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
            Assembly: GCF_000001735.4
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
            Camelineae; Arabidopsis.
REFERENCE   1  (bases 1 to 2201)
  AUTHORS   Lin,X., Kaul,S., Rounsley,S., Shea,T.P., Benito,M.I., Town,C.D.,
            Fujii,C.Y., Mason,T., Bowman,C.L., Barnstead,M., Feldblyum,T.V.,
```

Change region shown

○ Whole sequence (abbreviated view)
● Selected region
from: 12368220 to: 12370420

Customize view

○ Abbreviated view
● Customize

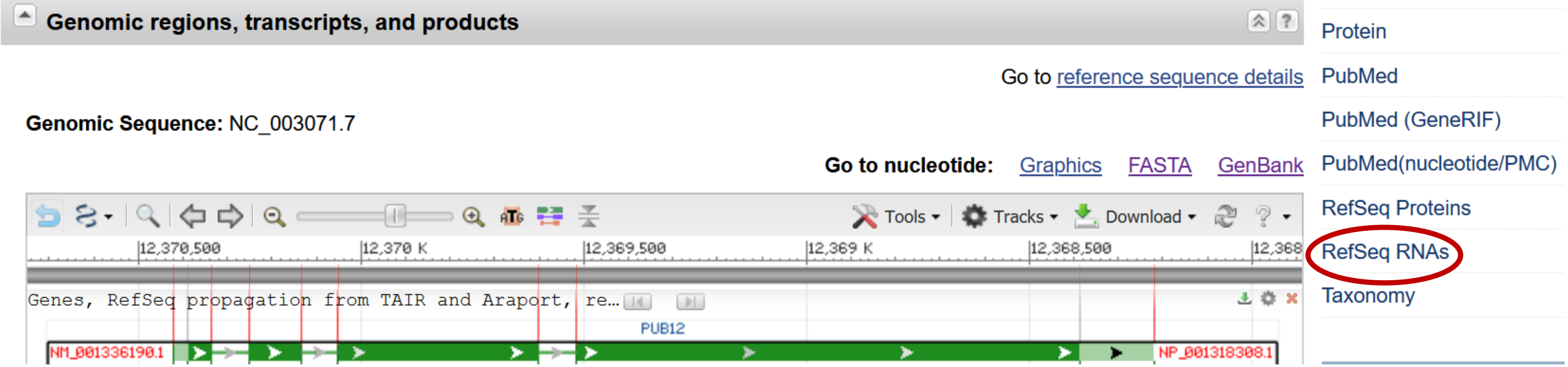
Basic Features
○ All features
● Gene, RNA, and CDS features only

Display options
☑ Show sequence
☑ Show reverse complement
□ Show gap features

https://www.ncbi.nlm.nih.gov/nuccore/NC_003071.7?report=genbank&from=12368220&to=12370420&strand=true

Đi đến trang protein của gen này

- Click on the RefSeq RNAs link in the “Related information” panel on the right.
- This takes you to the mRNA that encodes the protein you have been looking at (we are accessing the same record you accessed in Step 10 of the first part of the lab).



Genomic regions, transcripts, and products

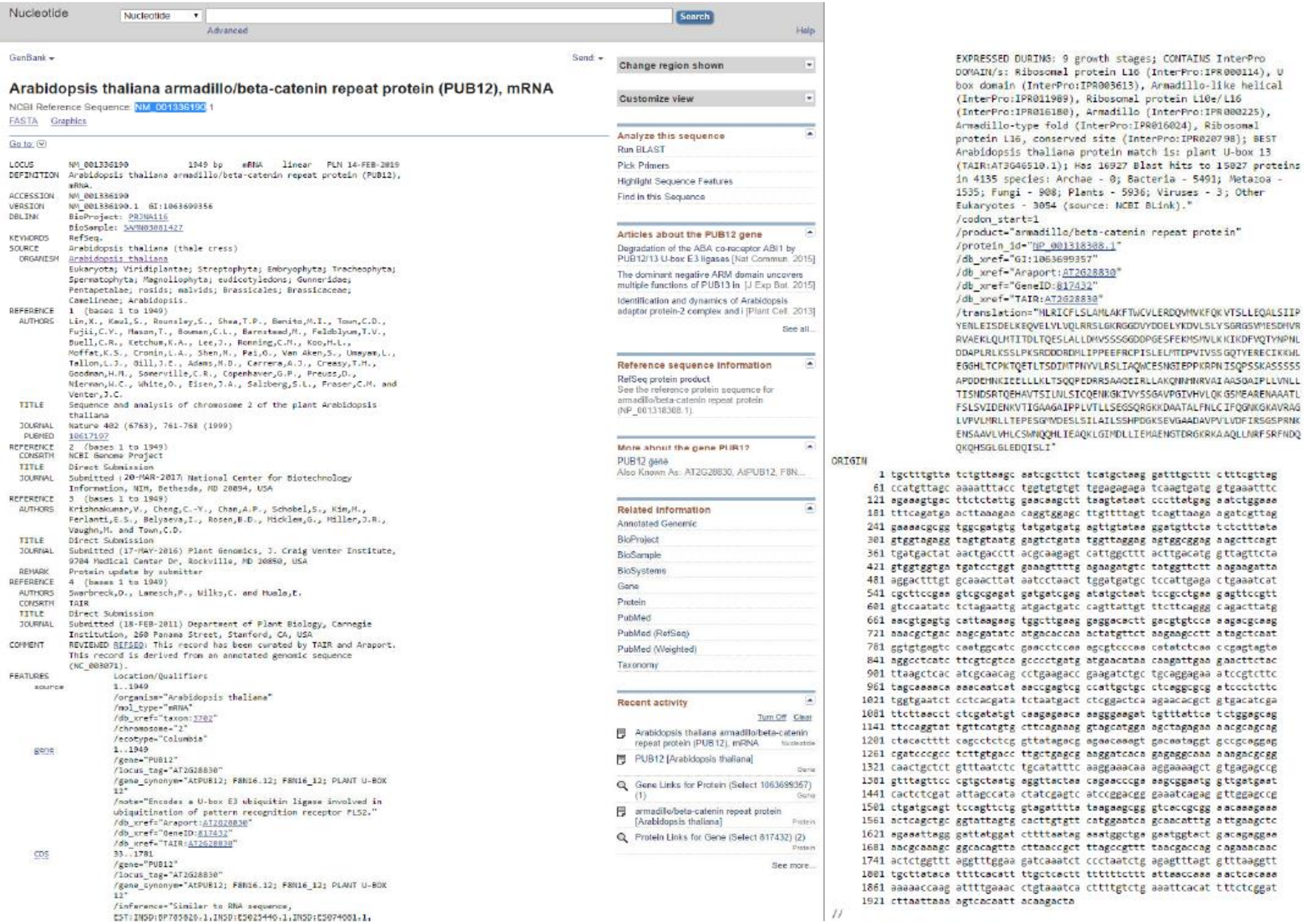
Go to reference sequence details

Genomic Sequence: NC_003071.7

Go to nucleotide: Graphics FASTA GenBank

Genes, RefSeq propagation from TAIR and Araport, re…

https://www.ncbi.nlm.nih.gov/gene/817432/



Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA

NCBI Reference Sequence: NM_001336190.1

FASTA Graphics

```
LOCUS       NM_001336190            1949 bp    mRNA    linear   PLN 14-FEB-2019
DEFINITION  Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12),
            mRNA.
ACCESSION   NM_001336190
VERSION     NM_001336190.1  GI:1063699356
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
            Camelineae; Arabidopsis.
REFERENCE   1  (bases 1 to 1949)
  AUTHORS   Lin,X., Kaul,S., Rounsley,S., Shea,T.P., Benito,M.I., Town,C.D.,
            Fujii,C.Y., Mason,T., Bowman,C.L., Barnstead,M., Feldblyum,T.V.,
            Buell,C.R., Ketchum,K.A., Lee,J., Ronning,C.M., Koo,H.L.,
            Moffat,K.S., Cronin,L.A., Shen,M., Pai,G., Van Aken,S., Umayam,L.,
            Tallon,L.J., Gill,J.E., Adams,M.D., Carrera,A.J., Creasy,T.H.,
            Goodman,H.M., Somerville,C.R., Copenhaver,G.P., Preuss,D.,
            Nierman,W.C., White,O., Eisen,J.A., Salzberg,S.L., Fraser,C.M. and
            Venter,J.C.
  TITLE     Sequence and analysis of chromosome 2 of the plant Arabidopsis
            thaliana
  JOURNAL   Nature 402 (6763), 761-768 (1999)
   PUBMED   10617197
REFERENCE   2  (bases 1 to 1949)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (20-MAR-2017) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (bases 1 to 1949)
  AUTHORS   Krishnakumar,V., Cheng,C.-Y., Chan,A.P., Schobel,S., Kim,M.,
            Ferlanti,E.S., Belyaeva,I., Rosen,B.D., Micklem,G., Miller,J.R.,
            Vaughn,M. and Town,C.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-MAY-2016) Plant Genomics, J. Craig Venter Institute,
            9704 Medical Center Dr, Rockville, MD 20850, USA
  REMARK    Protein update by submitter
REFERENCE   4  (bases 1 to 1949)
  AUTHORS   Swarbreck,D., Lamesch,P., Wilks,C. and Huala,E.
  CONSRTM   TAIR
  TITLE     Direct Submission
  JOURNAL   Submitted (18-FEB-2011) Department of Plant Biology, Carnegie
            Institution, 260 Panama Street, Stanford, CA, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by TAIR and Araport.
            This record is derived from an annotated genomic sequence
            (NC_003071).
FEATURES             Location/Qualifiers
     source          1..1949
                     /organism="Arabidopsis thaliana"
                     /mol_type="mRNA"
                     /db_xref="taxon:3702"
                     /chromosome="2"
                     /ecotype="Columbia"
     gene            1..1949
                     /gene="PUB12"
                     /locus_tag="AT2G28830"
                     /gene_synonym="AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX
                     12"
                     /note="Encodes a U-box E3 ubiquitin ligase involved in
                     ubiquitination of pattern recognition receptor FLS2."
                     /db_xref="Araport:AT2G28830"
                     /db_xref="GeneID:817432"
                     /db_xref="TAIR:AT2G28830"
     CDS             33..1781
                     /gene="PUB12"
                     /locus_tag="AT2G28830"
                     /gene_synonym="AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX
                     12"
                     /inference="Similar to RNA sequence,
                     EST:INSD:BP785820.1,INSD:ES025440.1,INSD:ES074081.1,
```

```
                     EXPRESSED DURING: 9 growth stages; CONTAINS InterPro
                     DOMAIN/s: Ribosomal protein L16 (InterPro:IPR000114), U
                     box domain (InterPro:IPR003613), Armadillo-like helical
                     (InterPro:IPR011989), Ribosomal protein L10e/L16
                     (InterPro:IPR016180), Armadillo (InterPro:IPR000225),
                     Armadillo-type fold (InterPro:IPR016024), Ribosomal
                     protein L16, conserved site (InterPro:IPR020798); BEST
                     Arabidopsis thaliana protein match is: plant U-box 13
                     (TAIR:AT3G46510.1); Has 16927 Blast hits to 15027 proteins
                     in 4135 species: Archae - 0; Bacteria - 5491; Metazoa -
                     1535; Fungi - 908; Plants - 5936; Viruses - 3; Other
                     Eukaryotes - 3054 (source: NCBI BLink)."
                     /codon_start=1
                     /product="armadillo/beta-catenin repeat protein"
                     /protein_id="NP_001318308.1"
                     /db_xref="GI:1063699357"
                     /db_xref="Araport:AT2G28830"
                     /db_xref="GeneID:817432"
                     /db_xref="TAIR:AT2G28830"
                     /translation="MLRICFLSLAMLAKFTWCVLERDQVMVKFQKVTSLLEQALSIIP
                     YENLEISDELKEQVELVLVQLRRSLGKRGGDVYDDELYKDVLSLYSGRGSVMESDMVR
                     RVAEKLQLMTITDLTQESLALLDMVSSSGGDDPGESFEKMSMVLKKIKDFVQTYNPNL
                     DDAPLRLKSSLPKSRDDDRDMLIPPEEFRCPISLELMTDPVIVSSGQTYERECIKKWL
                     EGGHLTCPKTQETLTSDIMTPNYVLRSLIAQWCESNGIEPPKRPNISQPSSKASSSSS
                     APDDEHNKIEELLLKLTSQQPEDRRSAAGEIRLLAKQNNHNRVAIAASGAIPLLVNLL
                     TISNDSRTQEHAVTSILNLSICQENKGKIVYSSGAVPGIVHVLQKGSMEARENAAATL
                     FSLSVIDENKVTIGAAGAIPPLVTLLSEGSQRGKKDAATALFNLCIFQGNKGKAVRAG
                     LVPVLMRLLTEPESGMVDESLSILAILSSHPDGKSEVGAADAVPVLVDFIRSGSPRNK
                     ENSAAVLVHLCSWNQQHLIEAQKLGIMDLLIEMAENGTDRGKRKAAQLLNRFSRFNDQ
                     QKQHSGLGLEDQISLI"
ORIGIN
        1 tgctttgtta tctgttaagc aatcgcttct tcatgctaag gatttgcttt ctttcgttag
       61 ccatgttagc aaaatttacc tggtgtgtgt tggagagaga tcaagtgatg gtgaaatttc
      121 agaaagtgac ttctctattg gaacaagctt taagtataat cccttatgag aatctggaaa
      181 tttcagatga acttaaagaa caggtggagc ttgttttagt tcagttaaga agatcgttag
      241 gaaaacgcgg tggcgatgtg tatgatgatg agttgtataa ggatgttcta tctctttata
      301 gtggtagagg tagtgtaatg gagtctgata tggttaggag agtggcggag aagcttcagt
      361 tgatgactat aactgacctt acgcaagagt cattggcttt acttgacatg gttagttcta
      421 gtggtggtga tgatcctggt gaaagttttg agaagatgtc tatggttctt aagaagatta
      481 aggactttgt gcaaacttat aatcctaact tggatgatgc tccattgaga ctgaaatcat
      541 cgcttccgaa gtcgcgagat gatgatcgag atatgctaat tccgcctgaa gagttccgtt
      601 gtccaatatc tctagaattg atgactgatc cagttattgt ttcttcaggg cagacttatg
      661 aacgtgagtg cattaagaag tggcttgaag gaggacactt gacgtgtcca aagacgcaag
      721 aaacgctgac aagcgatatc atgacaccaa actatgttct aagaagcctt atagctcaat
      781 ggtgtgagtc caatggcatc gaacctccaa agcgtcccaa catatctcaa ccgagtagta
      841 aggcctcatc ttcgtcgtca gcccctgatg atgaacataa caagattgaa gaacttctac
      901 ttaagctcac atcgcaacag cctgaagacc gaagatctgc tgcaggagaa atccgtcttc
      961 tagcaaaaca aaacaatcat aaccgagtcg ccattgctgc ctcaggcgcg atccctcttc
     1021 tggtgaatct cctcacgata tctaatgact ctcggactca agaacacgct gtgacatcga
     1081 ttcttaacct ctcgatatgt caagagaaca aagggaagat tgtttattca tctggagcag
     1141 ttccaggtat tgttcatgtg cttcagaaag gtagcatgga agctagagaa aacgcagcag
     1201 ctacactttt cagcctctcg gttatagacg agaacaaagt gacaataggt gccgcaggag
     1261 cgatcccgcc tcttgtgacc ttgctgagcg aaggatcaca gagaggcaaa aaagacgcgg
     1321 caactgctct gtttaatctc tgcatatttc aaggaaacaa aggaaaagct gtgagagccg
     1381 gtttagttcc cgtgctaatg aggttactaa cagaacccga aagcggaatg gttgatgaat
     1441 cactctcgat attagccata ctatcgagtc atccggacgg gaaatcagag gttggagccg
     1501 ctgatgcagt tccagttctg gtagatttta taagaagcgg gtcaccgcgg aacaaagaaa
     1561 actcagctgc ggtattagtg cacttgtgtt catggaatca gcaacatttg attgaagctc
     1621 agaaattagg gattatggat cttttaatag aaatggctga gaatggtact gacagaggaa
     1681 aacgcaaagc ggcacagtta cttaaccgct ttagccgttt taacgaccag cagaaacaac
     1741 actctggttt aggtttggaa gatcaaatct ccctaatctg agagtttagt gtttaaggtt
     1801 tgcttataca ttttcacatt ttgctcactt ttttttcttt attaaccaaa aactcacaaa
     1861 aaaaaccaag atttgaaac ctgtaaatca cttttgtctg aaattcacat tttctcggat
     1921 cttaattaaa agtcacaatt acaagacta
//
```

Analyze this sequence
- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

Articles about the PUB12 gene
- Degradation of the ABA co-receptor ABI1 by PUB12/13 U-box E3 ligases [Nat Commun. 2015]
- The dominant negative ARM domain uncovers multiple functions of PUB13 in [J Exp Bot. 2015]
- Identification and dynamics of Arabidopsis adaptor protein-2 complex and i [Plant Cell. 2013]
- See all...

Reference sequence information
- RefSeq protein product
  See the reference protein sequence for armadillo/beta-catenin repeat protein (NP_001318308.1).

More about the gene PUB12
- PUB12 gene
  Also Known As: AT2G28830, AtPUB12, F8N...

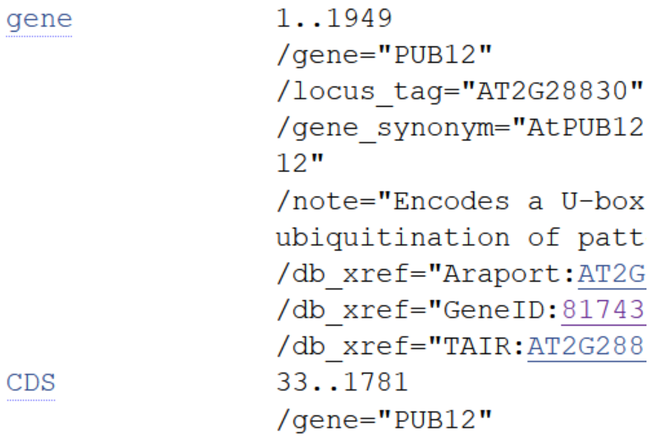
Related information
- Annotated Genomic
- BioProject
- BioSample
- BioSystems
- Gene
- Protein
- PubMed
- PubMed (RefSeq)
- PubMed (Weighted)
- Taxonomy

Recent activity
- Turn Off Clear
- Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA — Nucleotide
- PUB12 [Arabidopsis thaliana] — Gene
- Gene Links for Protein (Select 1063699357) (1) — Gene
- armadillo/beta-catenin repeat protein [Arabidopsis thaliana] — Protein
- Protein Links for Gene (Select 817432) (2) — Protein
- See more...

https://www.ncbi.nlm.nih.gov/nuccore/1063699356

https://www.ncbi.nlm.nih.gov/nuccore/1063699356

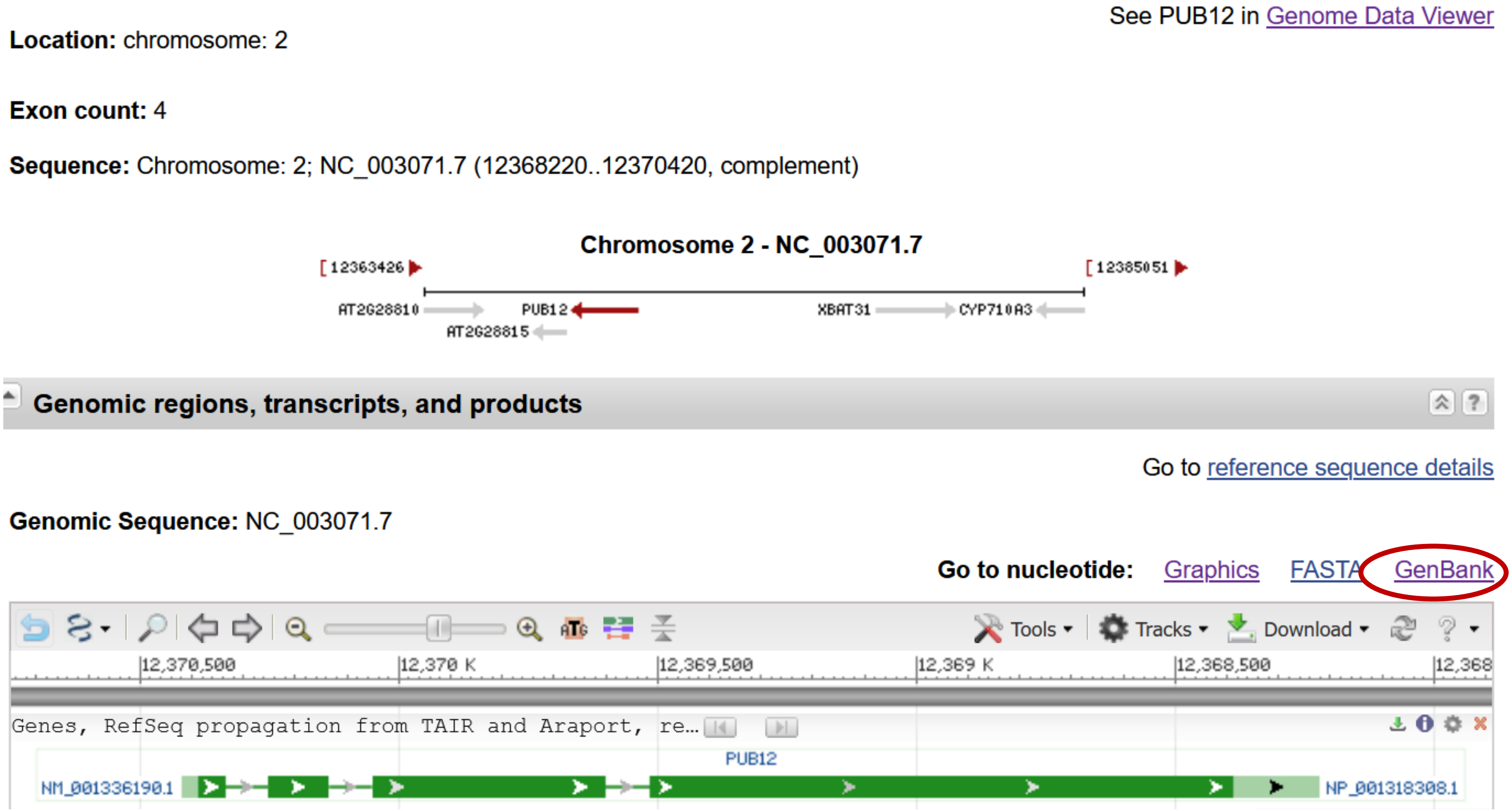
- Notice the feature list in the record.
- One Feature in the GenBank record is gene, and corresponds to base position 1 – 1949 on this record.
- Another features is the coding sequence (CDS), which corresponds to base position 33 – 1781.
- ***a. Given your biology background knowledge, why do you think these are different?***



```
gene                1..1949
                    /gene="PUB12"
                    /locus_tag="AT2G28830"
                    /gene_synonym="AtPUB12
                    12"
                    /note="Encodes a U-box
                    ubiquitination of patt
                    /db_xref="Araport:AT2G
                    /db_xref="GeneID:81743
                    /db_xref="TAIR:AT2G288
CDS                 33..1781
                    /gene="PUB12"
```

- Gene: 2201 nu, ở sợi bù (complement), do đó chạy từ chỉ số lớn về chỉ số nhỏ (12368220..12370420)
- mRNA: 1949 nu (do đã bị loại bỏ các đoạn introns) : 1..86,170..286, 370..819, 906..2201
- CDS (Coding Sequence):
  - Loại bỏ phần đầu và phần cuối (UTR: UnTranslated Region): phần không mã hóa amino acids
  - từ 33 – 1781, như vậy có 1749 nu => mã hóa cho 583 amino acid
- Trong đó codon đầu tiên là AUG (mã hóa cho Met), vẫn được tích hợp vào chuỗi amino acid.
- Codon cuối cùng là TGA, là codon kết thúc (stop), không mã hóa cho amino acid nào, và vì vậy protein này chỉ có 582 amino acids

Above the Sequence Viewer panel, click on the “Go to nucleotide: Genbank” link



See PUB12 in Genome Data Viewer

Location: chromosome: 2

Exon count: 4

Sequence: Chromosome: 2; NC_003071.7 (12368220..12370420, complement)

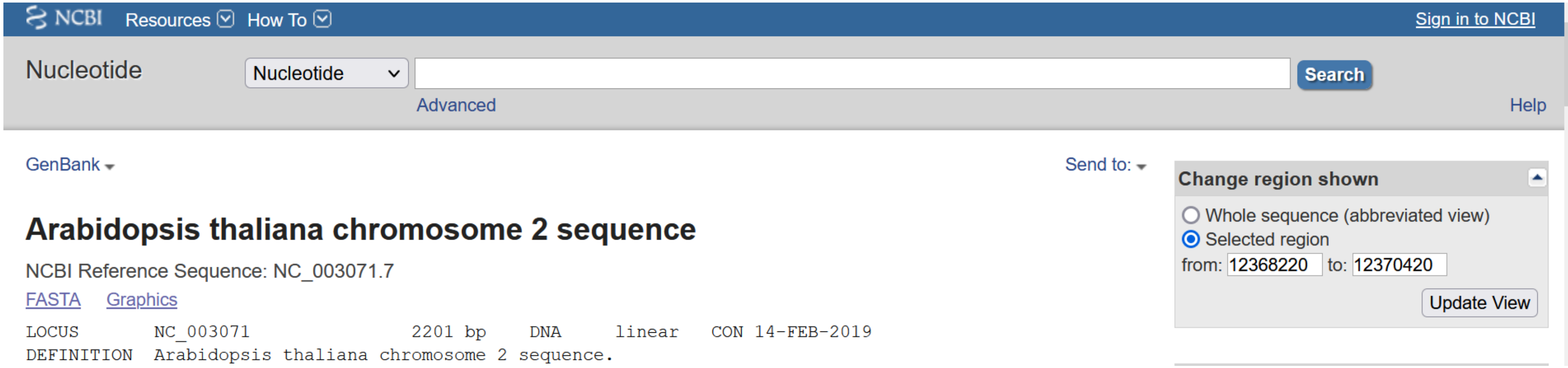
Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_003071.7

Go to nucleotide: Graphics FASTA GenBank

https://www.ncbi.nlm.nih.gov/gene/817432/



Arabidopsis thaliana chromosome 2 sequence

NCBI Reference Sequence: NC_003071.7

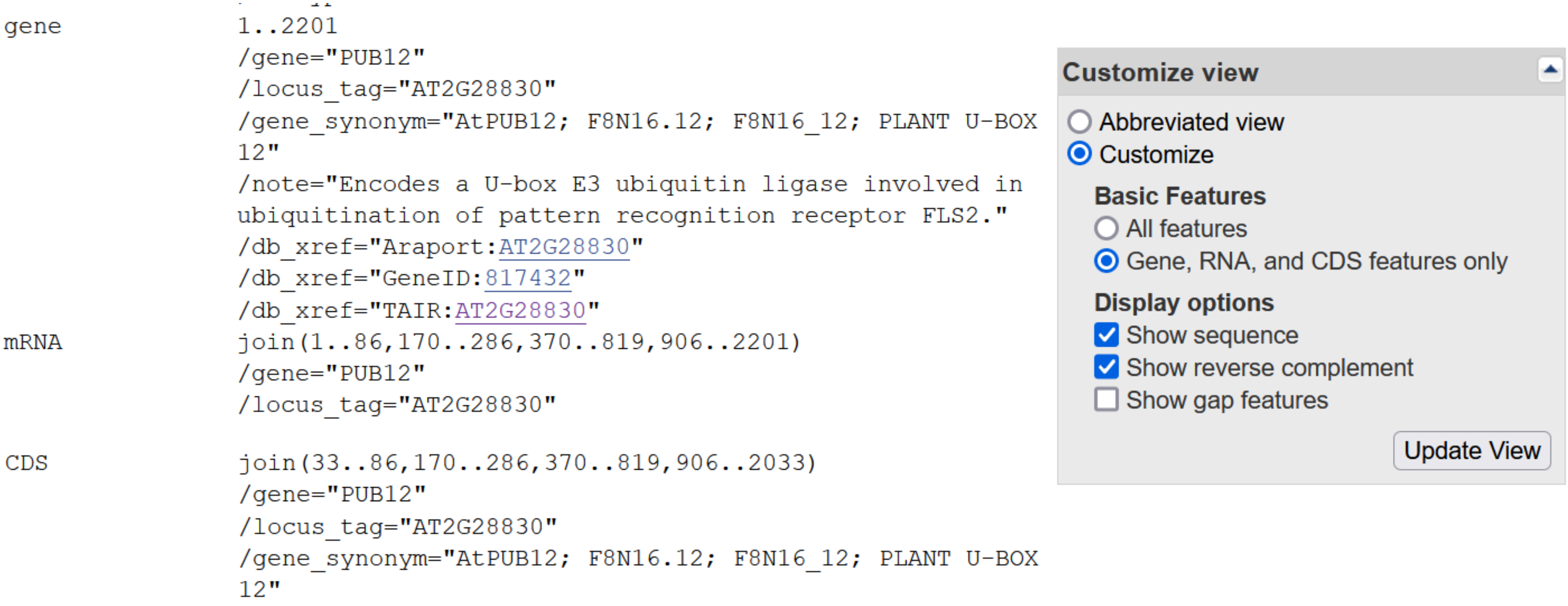
FASTA Graphics

```
LOCUS       NC_003071                2201 bp    DNA     linear   CON 14-FEB-2019
DEFINITION  Arabidopsis thaliana chromosome 2 sequence.
```

Change region shown

○ Whole sequence (abbreviated view)
● Selected region
from: 12368220 to: 12370420

Update View



```
gene            1..2201
                /gene="PUB12"
                /locus_tag="AT2G28830"
                /gene_synonym="AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX
                12"
                /note="Encodes a U-box E3 ubiquitin ligase involved in
                ubiquitination of pattern recognition receptor FLS2."
                /db_xref="Araport:AT2G28830"
                /db_xref="GeneID:817432"
                /db_xref="TAIR:AT2G28830"
mRNA            join(1..86,170..286,370..819,906..2201)
                /gene="PUB12"
                /locus_tag="AT2G28830"

CDS             join(33..86,170..286,370..819,906..2033)
                /gene="PUB12"
                /locus_tag="AT2G28830"
                /gene_synonym="AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX
                12"
```

Customize view

○ Abbreviated view
● Customize

Basic Features

○ All features
● Gene, RNA, and CDS features only

Display options

☑ Show sequence
☑ Show reverse complement
□ Show gap features

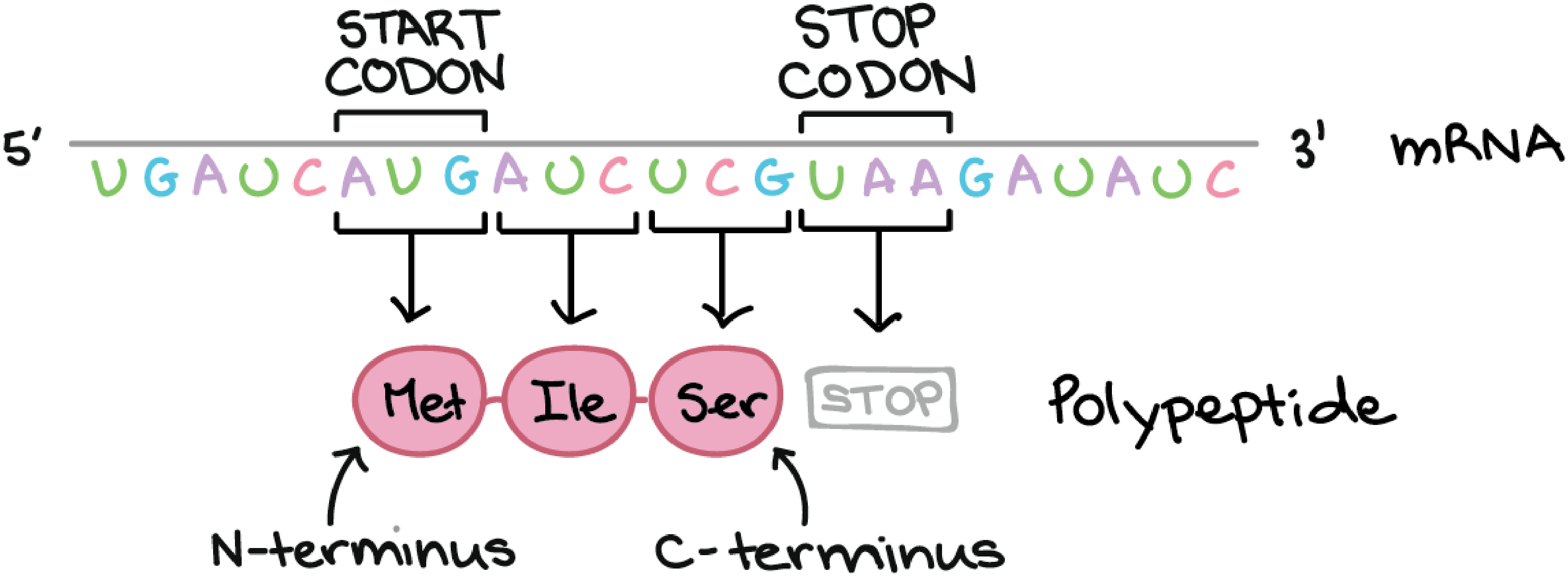
Update View

https://www.ncbi.nlm.nih.gov/nuccore/NC_003071.7?report=genbank&from=12368220&to=12370420&strand=true

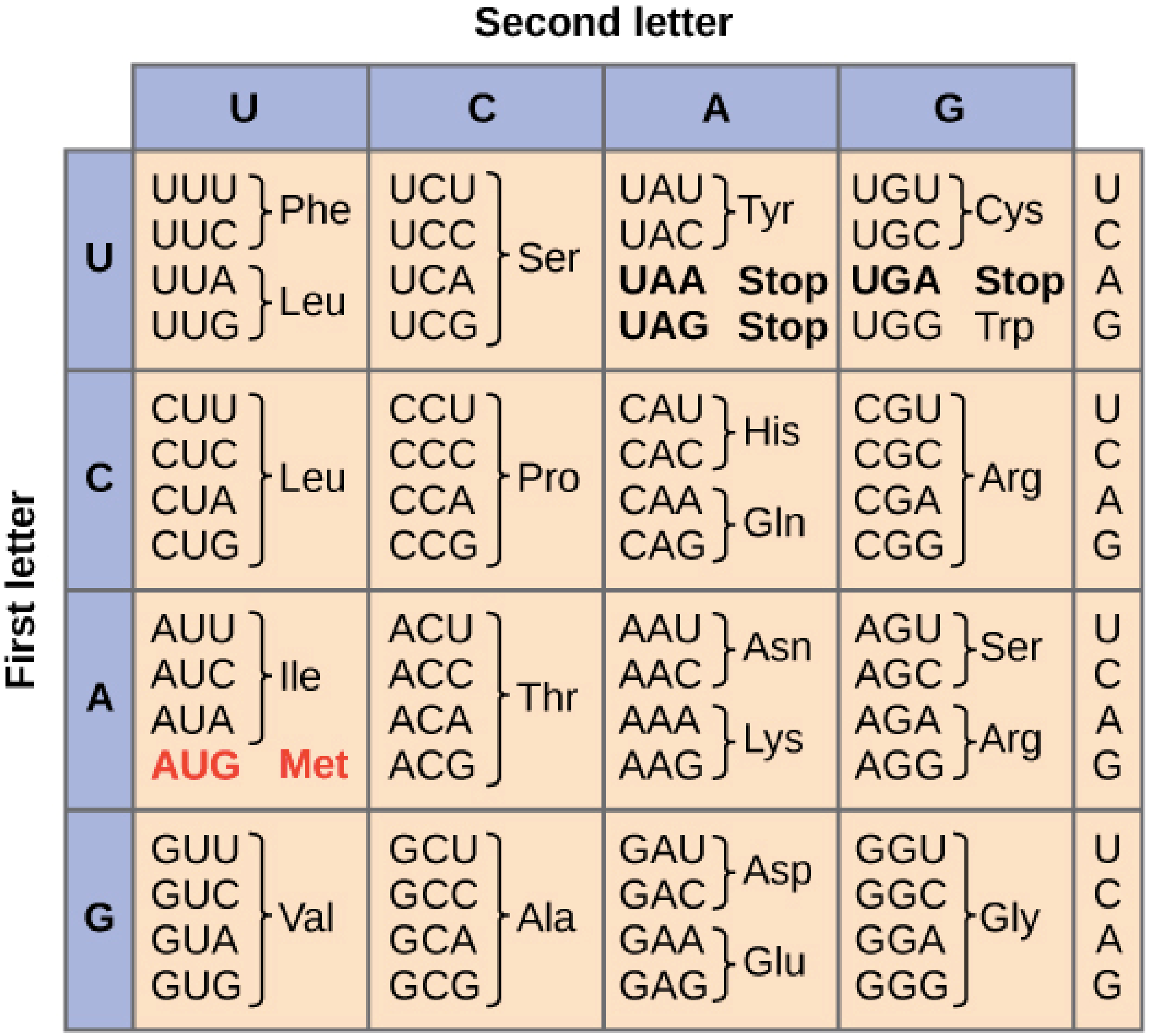
- You will be taken you to the genomic region that encodes the mRNA you were just looking at.
- Notice how the gene feature corresponds to positions 1–2201, while the mRNA feature corresponds to positions 1–86, 170–286, 370–819, and 906–2201 and the CDS feature corresponds to nucleotide positions 33–86, 170–286, 370–819, and 906–2033.
- You may have remarked that the sequence from the chromosome has been reverse complemented.
- ***b. Again, why are these different? Tip: recall the Central Dogma of Molecular Biology!***

https://www.ncbi.nlm.nih.gov/nuccore/NC_003071.7?report=genbank&from=12368220&to=12370420&strand=true

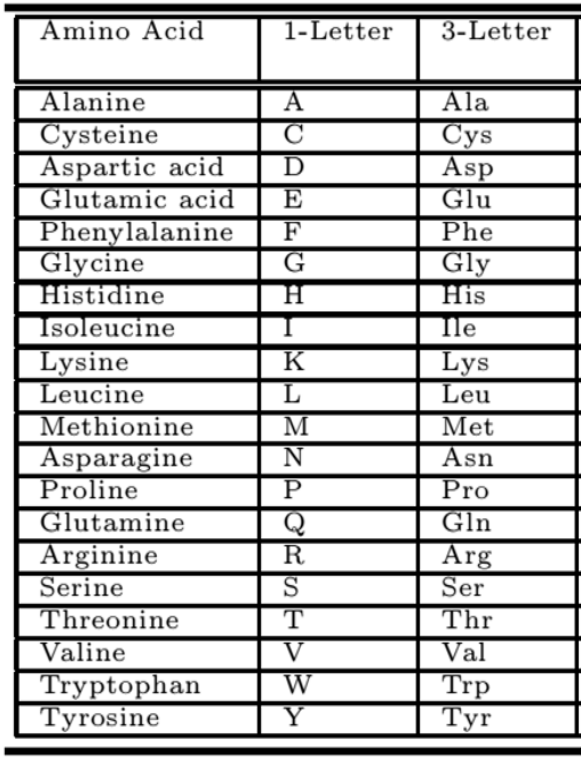
Central Dogma of Molecular Biology



Genetic code



First letter	Second letter								
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		**UAA**	**Stop**	**UGA**	**Stop**	A
	UUG		UCG		**UAG**	**Stop**	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

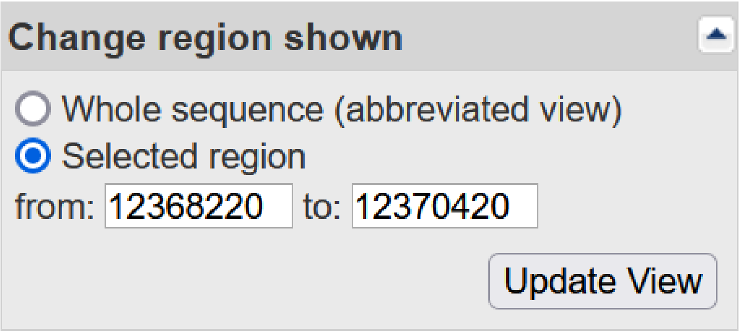


Amino Acid	1-Letter	3-Letter
Alanine	A	Ala
Cysteine	C	Cys
Aspartic acid	D	Asp
Glutamic acid	E	Glu
Phenylalanine	F	Phe
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Lysine	K	Lys
Leucine	L	Leu
Methionine	M	Met
Asparagine	N	Asn
Proline	P	Pro
Glutamine	Q	Gln
Arginine	R	Arg
Serine	S	Ser
Threonine	T	Thr
Valine	V	Val
Tryptophan	W	Trp
Tyrosine	Y	Tyr

- View Reverse Complement: Notice how the gene feature corresponds to positions 1–2201,
    - while the mRNA feature corresponds to positions 1–86, 170–286, 370–819, and 906–2201 and
    - the CDS feature corresponds to nucleotide positions 33–86, 170–286, 370–819, and 906–2033.
- View bình thường:
    - mRNA: complement(join(1..1296, 1383..1832, 1916..2032, 2116..2201))
    - CDS: complement(join(169..1296, 1383..1832, 1916..2032, 2116..2169))

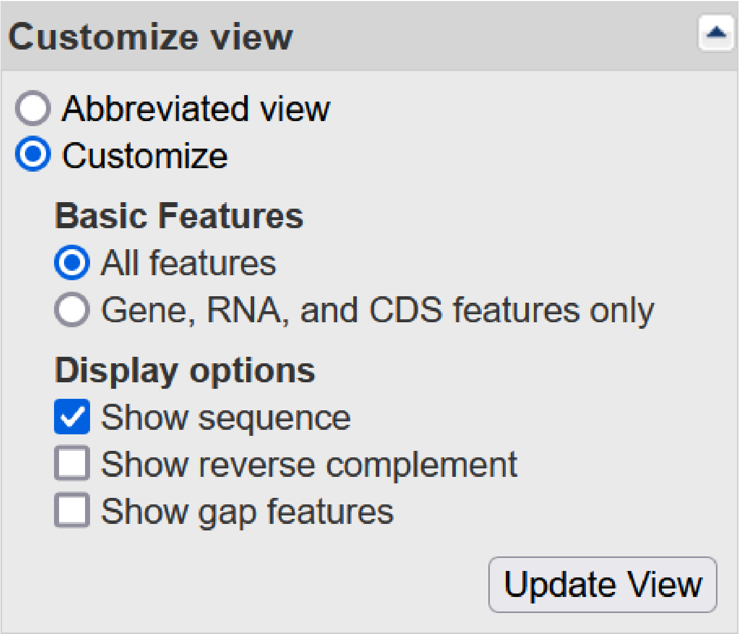
Xem ở chế độ bình thường

gene : complement(1..2201)
mRNA: complement(join(1..1296,
1383..1832, 1916..2032,
2116..2201))
CDS: complement(join(169..1296,
1383..1832, 1916..2032,
2116..2169))
/translation="MLRI…"
ORIGIN
2161 ccttag***cat***g aagaagcgat
tgcttaacag ataacaaagc a



Change region shown

○ Whole sequence (abbreviated view)
● Selected region
from: 12368220 to: 12370420



Customize view

○ Abbreviated view
● Customize

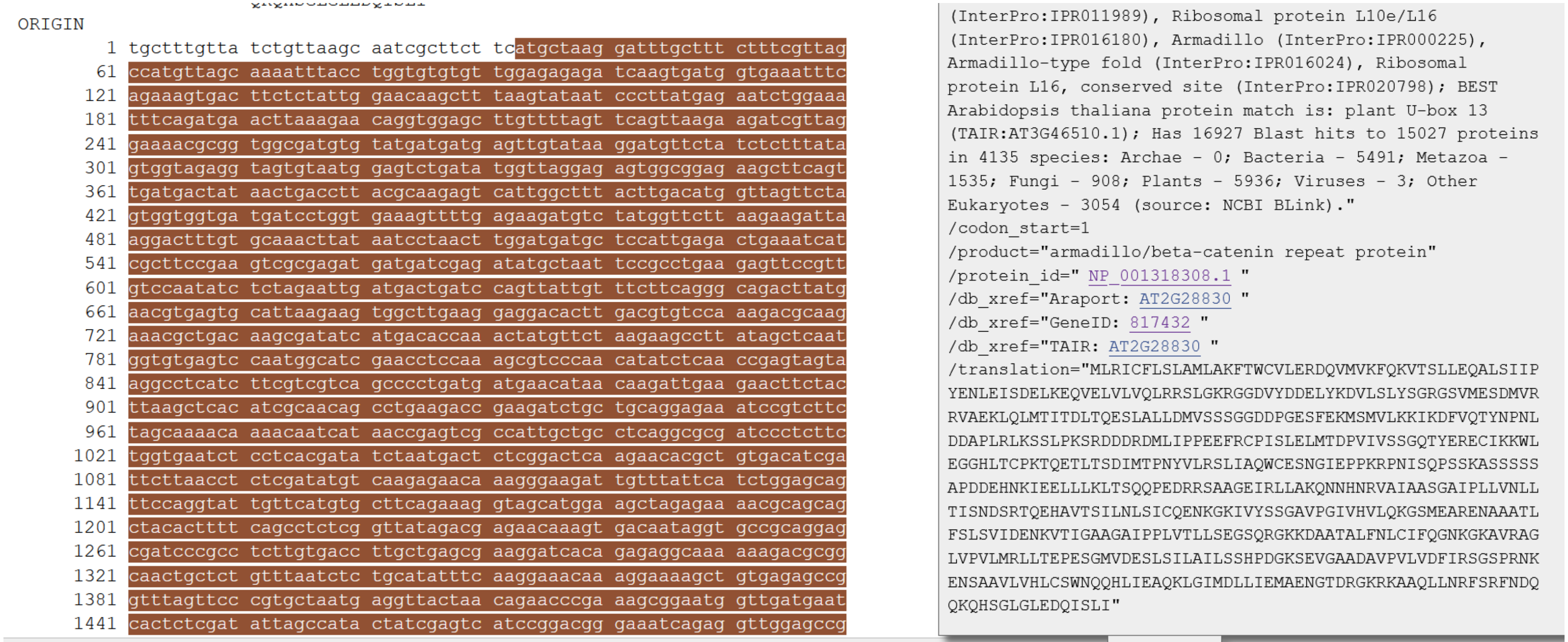
Basic Features
● All features
○ Gene, RNA, and CDS features only

Display options
☑ Show sequence
□ Show reverse complement
□ Show gap features

https://www.ncbi.nlm.nih.gov/nuccore/NC_003071.7?report=genbank&from=12368220&to=12370420&strand=true

- Let's return the mRNA record we were previously working with (NM_001336190).
- Click on the CDS link.
- Now you are looking at the information for the coding sequence, as opposed to the whole gene or protein (highlighted in brown).
- Using the “Display: FASTA” option in the grey bar at the bottom of the page generate a FASTA-formatted version of the CDS.

https://www.ncbi.nlm.nih.gov/nuccore/1063699356

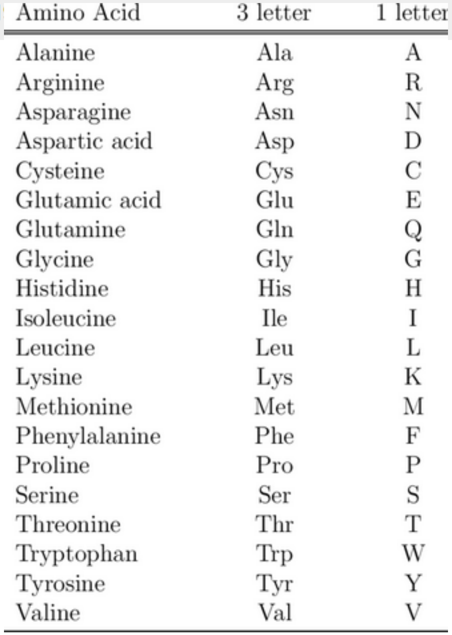


```
                    QKQHSGLGLEDQISLI
ORIGIN
        1 tgctttgtta tctgttaagc aatcgcttct tcatgctaag gatttgcttt ctttcgttag
       61 ccatgttagc aaaatttacc tggtgtgtgt tggagagaga tcaagtgatg gtgaaatttc
      121 agaaagtgac ttctctattg gaacaagctt taagtataat cccttatgag aatctggaaa
      181 tttcagatga acttaaagaa caggtggagc ttgttttagt tcagttaaga agatcgttag
      241 gaaaacgcgg tggcgatgtg tatgatgatg agttgtataa ggatgttcta tctctttata
      301 gtggtagagg tagtgtaatg gagtctgata tggttaggag agtggcggag aagcttcagt
      361 tgatgactat aactgacctt acgcaagagt cattggcttt acttgacatg gttagttcta
      421 gtggtggtga tgatcctggt gaaagttttg agaagatgtc tatggttctt aagaagatta
      481 aggactttgt gcaaacttat aatcctaact tggatgatgc tccattgaga ctgaaatcat
      541 cgcttccgaa gtcgcgagat gatgatcgag atatgctaat tccgcctgaa gagttccgtt
      601 gtccaatatc tctagaattg atgactgatc cagttattgt ttcttcaggg cagacttatg
      661 aacgtgagtg cattaagaag tggcttgaag gaggacactt gacgtgtcca aagacgcaag
      721 aaacgctgac aagcgatatc atgacaccaa actatgttct aagaagcctt atagctcaat
      781 ggtgtgagtc caatggcatc gaacctccaa agcgtcccaa catatctcaa ccgagtagta
      841 aggcctcatc ttcgtcgtca gcccctgatg atgaacataa caagattgaa gaacttctac
      901 ttaagctcac atcgcaacag cctgaagacc gaagatctgc tgcaggagaa atccgtcttc
      961 tagcaaaaca aaacaatcat aaccgagtcg ccattgctgc ctcaggcgcg atccctcttc
     1021 tggtgaatct cctcacgata tctaatgact ctcggactca agaacacgct gtgacatcga
     1081 ttcttaacct ctcgatatgt caagagaaca aagggaagat tgtttattca tctggagcag
     1141 ttccaggtat tgttcatgtg cttcagaaag gtagcatgga agctagagaa aacgcagcag
     1201 ctacactttt cagcctctcg gttatagacg agaacaaagt gacaataggt gccgcaggag
     1261 cgatcccgcc tcttgtgacc ttgctgagcg aaggatcaca gagaggcaaa aaagacgcgg
     1321 caactgctct gtttaatctc tgcatatttc aaggaaacaa aggaaaagct gtgagagccg
     1381 gtttagttcc cgtgctaatg aggttactaa cagaacccga aagcggaatg gttgatgaat
     1441 cactctcgat attagccata ctatcgagtc atccggacgg gaaatcagag gttggagccg
```

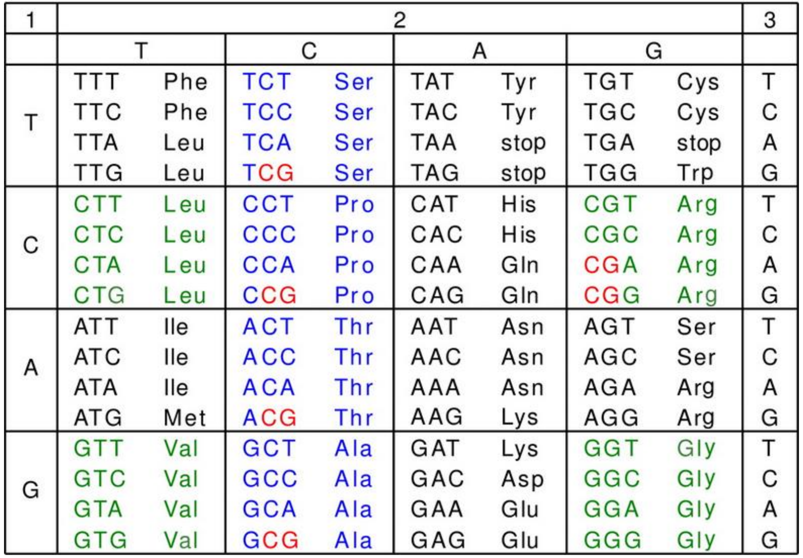
```
(InterPro:IPR011989), Ribosomal protein L10e/L16
(InterPro:IPR016180), Armadillo (InterPro:IPR000225),
Armadillo-type fold (InterPro:IPR016024), Ribosomal
protein L16, conserved site (InterPro:IPR020798); BEST
Arabidopsis thaliana protein match is: plant U-box 13
(TAIR:AT3G46510.1); Has 16927 Blast hits to 15027 proteins
in 4135 species: Archae - 0; Bacteria - 5491; Metazoa -
1535; Fungi - 908; Plants - 5936; Viruses - 3; Other
Eukaryotes - 3054 (source: NCBI BLink)."
/codon_start=1
/product="armadillo/beta-catenin repeat protein"
/protein_id=" NP_001318308.1 "
/db_xref="Araport: AT2G28830 "
/db_xref="GeneID: 817432 "
/db_xref="TAIR: AT2G28830 "
/translation="MLRICFLSLAMLAKFTWCVLERDQVMVKFQKVTSLLEQALSIIP
YENLEISDELKEQVELVLVQLRRSLGKRGGDVYDDELYKDVLSLYSGRGSVMESDMVR
RVAEKLQLMTITDLTQESLALLDMVSSSGGDDPGESFEKMSMVLKKIKDFVQTYNPNL
DDAPLRLKSSLPKSRDDDRDMLIPPEEFRCPISLELMTDPVIVSSGQTYERECIKKWL
EGGHLTCPKTQETLTSDIMTPNYVLRSLIAQWCESNGIEPPKRPNISQPSSKASSSSS
APDDEHNKIEELLLKLTSQQPEDRRSAAGEIRLLAKQNNHNRVAIAASGAIPLLVNLL
TISNDSRTQEHAVTSILNLSICQENKGKIVYSSGAVPGIVHVLQKGSMEARENAAATL
FSLSVIDENKVTIGAAGAIPPLVTLLSEGSQRGKKDAATALFNLCIFQGNKGKAVRAG
LVPVLMRLLTEPESGMVDESLSILAILSSHPDGKSEVGAADAVPVLVDFIRSGSPRNK
ENSAAVLVHLCSWNQQHLIEAQKLGIMDLLIEMAENGTDRGKRKAAQLLNRFSRFNDQ
QKQHSGLGLEDQISLI"
```

CDS ⌄ Feature « ‹ 1 of 1 › » NM_0013361 Details ⌄ Display: FASTA GenBank Help

- atg cta ag g att tgc ttt
- M L R I C F
- cta atc **tg a**
- LI <stop>

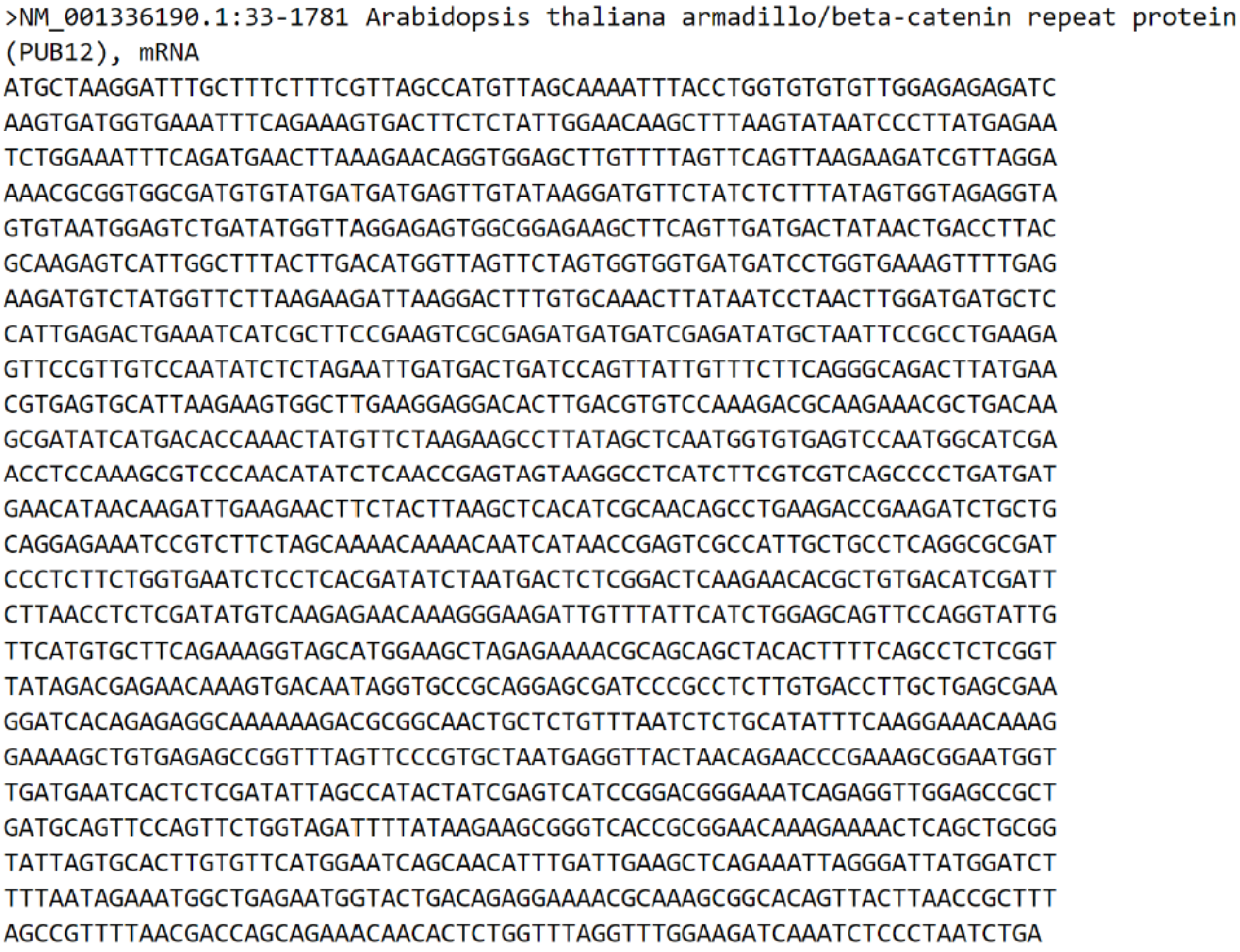


Amino Acid	3 letter	1 letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V



1	2								3
	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	C
	TTA	Leu	TCA	Ser	TAA	stop	TGA	stop	A
	TTG	Leu	TCG	Ser	TAG	stop	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	ATA	Ile	ACA	Thr	AAA	Asn	AGA	Arg	A
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GTT	Val	GCT	Ala	GAT	Lys	GGT	Gly	T
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Sequence in FASTA text format



```
>NM_001336190.1:33-1781 Arabidopsis thaliana armadillo/beta-catenin repeat protein
(PUB12), mRNA
ATGCTAAGGATTTGCTTTCTTTCGTTAGCCATGTTAGCAAAATTTACCTGGTGTGTGTTGGAGAGAGATC
AAGTGATGGTGAAATTTCAGAAAGTGACTTCTCTATTGGAACAAGCTTTAAGTATAATCCCTTATGAGAA
TCTGGAAATTTCAGATGAACTTAAAGAACAGGTGGAGCTTGTTTTAGTTCAGTTAAGAAGATCGTTAGGA
AAACGCGGTGGCGATGTGTATGATGATGAGTTGTATAAGGATGTTCTATCTCTTTATAGTGGTAGAGGTA
GTGTAATGGAGTCTGATATGGTTAGGAGAGTGGCGGAGAAGCTTCAGTTGATGACTATAACTGACCTTAC
GCAAGAGTCATTGGCTTTACTTGACATGGTTAGTTCTAGTGGTGGTGATGATCCTGGTGAAAGTTTTGAG
AAGATGTCTATGGTTCTTAAGAAGATTAAGGACTTTGTGCAAACTTATAATCCTAACTTGGATGATGCTC
CATTGAGACTGAAATCATCGCTTCCGAAGTCGCGAGATGATGATCGAGATATGCTAATTCCGCCTGAAGA
GTTCCGTTGTCCAATATCTCTAGAATTGATGACTGATCCAGTTATTGTTTCTTCAGGGCAGACTTATGAA
CGTGAGTGCATTAAGAAGTGGCTTGAAGGAGGACACTTGACGTGTCCAAAGACGCAAGAAACGCTGACAA
GCGATATCATGACACCAAACTATGTTCTAAGAAGCCTTATAGCTCAATGGTGTGAGTCCAATGGCATCGA
ACCTCCAAAGCGTCCCAACATATCTCAACCGAGTAGTAAGGCCTCATCTTCGTCGTCAGCCCCTGATGAT
GAACATAACAAGATTGAAGAACTTCTACTTAAGCTCACATCGCAACAGCCTGAAGACCGAAGATCTGCTG
CAGGAGAAATCCGTCTTCTAGCAAAACAAAACAATCATAACCGAGTCGCCATTGCTGCCTCAGGCGCGAT
CCCTCTTCTGGTGAATCTCCTCACGATATCTAATGACTCTCGGACTCAAGAACACGCTGTGACATCGATT
CTTAACCTCTCGATATGTCAAGAGAACAAAGGGAAGATTGTTTATTCATCTGGAGCAGTTCCAGGTATTG
TTCATGTGCTTCAGAAAGGTAGCATGGAAGCTAGAGAAAACGCAGCAGCTACACTTTTCAGCCTCTCGGT
TATAGACGAGAACAAAGTGACAATAGGTGCCGCAGGAGCGATCCCGCCTCTTGTGACCTTGCTGAGCGAA
GGATCACAGAGAGGCAAAAAAGACGCGGCAACTGCTCTGTTTAATCTCTGCATATTTCAAGGAAACAAAG
GAAAAGCTGTGAGAGCCGGTTTAGTTCCCGTGCTAATGAGGTTACTAACAGAACCCGAAAGCGGAATGGT
TGATGAATCACTCTCGATATTAGCCATACTATCGAGTCATCCGGACGGGAAATCAGAGGTTGGAGCCGCT
GATGCAGTTCCAGTTCTGGTAGATTTTATAAGAAGCGGGTCACCGCGGAACAAAGAAAACTCAGCTGCGG
TATTAGTGCACTTGTGTTCATGGAATCAGCAACATTTGATTGAAGCTCAGAAATTAGGGATTATGGATCT
TTTAATAGAAATGGCTGAGAATGGTACTGACAGAGGAAAACGCAAAGCGGCACAGTTACTTAACCGCTTT
AGCCGTTTTAACGACCAGCAGAAACAACACTCTGGTTTAGGTTTGGAAGATCAAATCTCCCTAATCTGA
```

https://www.ncbi.nlm.nih.gov/nuccore/NM_001336190.1?from=33&to=1781&report=fasta

```
Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA
NCBI Reference Sequence: NM_001336190.1
GenBank Graphics

>NM_001336190.1:33-1781 Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA
ATGCTAAGGATTTGCTTTCTTTCGTTAGCCATGTTAGCAAAATTTACCTGGTGTGTGTTGGAGAGAGATC
AAGTGATGGTGAAATTTCAGAAAGTGACTTCTCTATTGGAACAAGCTTTAAGTATAATCCCTTATGAGAA
TCTGGAAATTTCAGATGAACTTAAAGAACAGGTGGAGCTTGTTTTAGTTCAGTTAAGAAGATCGTTAGGA
AAACGCGGTGGCGATGTGTATGATGATGAGTTGTATAAGGATGTTCTATCTCTTTATAGTGGTAGAGGTA
GTGTAATGGAGTCTGATATGGTTAGGAGAGTGGCGGAGAAGCTTCAGTTGATGACTATAACTGACCTTAC
GCAAGAGTCATTGGCTTTACTTGACATGGTTAGTTCTAGTGGTGGTGATGATCCTGGTGAAAGTTTTGAG
AAGATGTCTATGGTTCTTAAGAAGATTAAGGACTTTGTGCAAACTTATAATCCTAACTTGGATGATGCTC
CATTGAGACTGAAATCATCGCTTCCGAAGTCGCGAGATGATGATCGAGATATGCTAATTCCGCCTGAAGA
GTTCCGTTGTCCAATATCTCTAGAATTGATGACTGATCCAGTTATTGTTTCTTCAGGGCAGACTTATGAA
CGTGAGTGCATTAAGAAGTGGCTTGAAGGAGGACACTTGACGTGTCCAAAGACGCAAGAAACGCTGACAA
GCGATATCATGACACCAAACTATGTTCTAAGAAGCCTTATAGCTCAATGGTGTGAGTCCAATGGCATCGA
ACCTCCAAAGCGTCCCAACATATCTCAACCGAGTAGTAAGGCCTCATCTTCGTCGTCAGCCCCTGATGAT
GAACATAACAAGATTGAAGAACTTCTACTTAAGCTCACATCGCAACAGCCTGAAGACCGAAGATCTGCTG
CAGGAGAAATCCGTCTTCTAGCAAAACAAAACAATCATAACCGAGTCGCCATTGCTGCCTCAGGCGCGAT
CCCTCTTCTGGTGAATCTCCTCACGATATCTAATGACTCTCGGACTCAAGAACACGCTGTGACATCGATT
CTTAACCTCTCGATATGTCAAGAGAACAAAGGGAAGATTGTTTATTCATCTGGAGCAGTTCCAGGTATTG
TTCATGTGCTTCAGAAAGGTAGCATGGAAGCTAGAGAAAACGCAGCAGCTACACTTTTCAGCCTCTCGGT
TATAGACGAGAACAAAGTGACAATAGGTGCCGCAGGAGCGATCCCGCCTCTTGTGACCTTGCTGAGCGAA
GGATCACAGAGAGGCAAAAAAGACGCGGCAACTGCTCTGTTTAATCTCTGCATATTTCAAGGAAACAAAG
GAAAAGCTGTGAGAGCCGGTTTAGTTCCCGTGCTAATGAGGTTACTAACAGAACCCGAAAGCGGAATGGT
TGATGAATCACTCTCGATATTAGCCATACTATCGAGTCATCCGGACGGGAAATCAGAGGTTGGAGCCGCT
GATGCAGTTCCAGTTCTGGTAGATTTTATAAGAAGCGGGTCACCGCGGAACAAAGAAAACTCAGCTGCGG
TATTAGTGCACTTGTGTTCATGGAATCAGCAACATTTGATTGAAGCTCAGAAATTAGGGATTATGGATCT
TTTAATAGAAATGGCTGAGAATGGTACTGACAGAGGAAAACGCAAAGCGGCACAGTTACTTAACCGCTTT
AGCCGTTTTAACGACCAGCAGAAACAACACTCTGGTTTAGGTTTGGAAGATCAAATCTCCCTAATCTGA
```

https://www.ncbi.nlm.nih.gov/nuccore/NM_001336190.1?from=33&to=1781&report=fasta