Genomic Database Conundrum:

Widespread Misannotation of rRNA Sequences as Protein Sequences

by

Miranda Raymond

A Senior Honors Project Presented to the

Honors College

East Carolina University

In Partial Fulfillment of the

Requirements for

Graduation with Honors

by

Miranda Raymond

Greenville, NC

December 2017

Approved by:

Dr. John Stiller

Department of Biology, Thomas Harriot College of Arts and Sciences

Abstract

The genomics revolution introduced affordable technology capable of rapidly analyzing and comparing massive amounts of biological sequence data. Using the Basic Local Alignment Search Tool (BLAST) program on the National Center for Biotechnology Information (NCBI) website, a highly expressed gene sequence obtained from the plant *Leptosiphon jepsonii* was analyzed. This sequence was compared against other sequences archived in the NCBI database for similarities. These comparisons encompassed various phyla of life including other green plants, fungi, metazoans, algae and single-celled organisms. The original sequence query was compared to inferred protein sequences. Then the mRNA sequences corresponding to these proteins were analyzed against complete nucleotide accessions through reciprocal BLAST searches to ensure accuracy of results. The most similar sequences from these reciprocal BLAST searches were rRNA rather than mRNA sequences. This result indicates that numerous accessions in NCBI are inappropriately characterized as mRNAs and proteins, rather than ribosomal sequences. To explore the breadth of this misannotation issue, sequences from a wide range of organisms, including model genomes, were also examined. This study indicates that rapid, automated computational analyses of massive amounts of sequence data, combined with a heightened focus on novel findings, has led to a sizable influx of erroneous data within even the most reputable databases.

Table of Contents

Abbreviations

DNA, deoxyribonucleic acid; NCBI, National Center for Biotechnology Information; BLAST, Basic Local Alignment Search Tool; mRNA, messenger ribonucleic acid; rRNA, ribosomal ribonucleic acid; RNAseq, massive parallel RNA sequencing; cDNA, complementary deoxyribonucleic acid; bp, DNA base pairs; ITS, internal transcribed spacers; ORF, open reading frame

Introduction

The human genome is composed of DNA, which encodes information necessary for the production and maintenance of life. For the information within DNA sequence to be useful to a cell, it must be transcribed into an RNA sequence, or transcript. A transcriptome is the collection of RNA transcripts within a cell. The size and function of transcriptomes vary depending on the organism, cell type and function. Transcriptome variation within the same organism results from different kinds and levels of expression of genes in different cell types (Transcriptome 2017).

RNA is present in multiple forms (Table 1). The prominent type is called mRNA and it is the message transcribed from genes that encode proteins. These mRNA transcripts are translated by the ribosome into amino acids, where they are assembled into chains of amino acids that make up proteins. The ribosomal core is made up of catalytic, structural rRNA molecules (Ribosomes, Transcription, and Translation 2017). The 18S, 5.8S, and 28S genes alternate with internal transcribed (ITS) regions that are removed before functional rRNAs are assembled into the main ribosomal subunit (see Figure 1). These functional regions are highly conserved across the tree of life and are easily characterized as rRNA.

Type of rRNA	Key Function
ribosomal RNA (rRNA)	structural and catalytic components of RNA enzymes, called ribozymes, that catalyze biochemical reactions
messenger RNA (mRNA)	translated into a protein by transfer RNA and the ribozyme

Table 1. Compares the functions of both rRNA and mRNA.

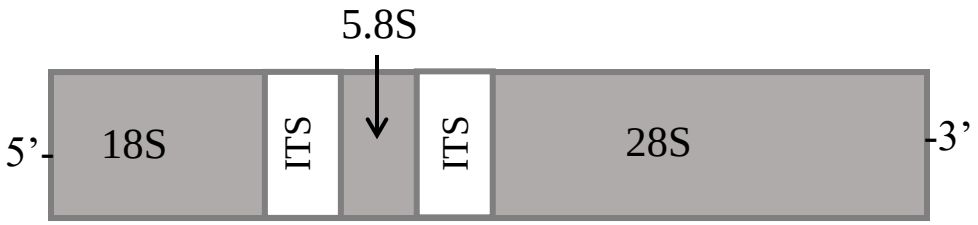


Figure 1. A diagram of the highly conserved regions of the rRNA genes and less conserved ITS regions (derived from Winnebeck et. al 2010)

The genomics revolution set in motion by the completion of the Human Genome Project in 2003 (Jenkins et al. 2005) led to rapid technological developments in the field of "transcriptomics." Transcriptomics is the study of the transcriptomes in cells and is useful in determining the functional elements of the genome. It aims at categorizing all transcripts by species and measuring the expression of each gene in different cell types and developmental conditions.

Next-generation sequencing has transformed the field of transcriptomics through its ability to process millions of sequence reads in parallel. Such high-throughput sequencing approaches allow scientists to rapidly map and quantify transcriptomes with relative ease, leading to an influx of novel mRNA sequence data into public databases (Mardis 2008). A popular method of next-generation sequencing is known as RNAseq, which analyzes a collection

of all mRNAs isolated from cells. The approach converts them to fragments of cDNA, then each molecule is sequenced to acquire short reads (between 30 and 400 bp) from one or both ends (Wang et al. 2009). The use of RNAseq has transformed modern genome sequencing procedures by detecting evidence for the expression of opening reading frames without stop codons, which could represent a gene. These particular ORFs can then be used to identify genes that encode proteins.

By examining the similarity of a newly discovered biological sequence to other known sequences, scientists can infer possible functions of the sequence in question. The NCBI website features the BLAST sequence similarity search, which performs local alignments of sequences within the database. A widely used tool in bioinformatics, BLAST, can be used to compare RNAseq data to the RefSeq public database of referenced annotations from all organisms. RefSeq is derived from the submissions to the redundant archival database GenBank; however, RefSeq theoretically provides an accurate, non-redundant annotation of nucleotide and protein sequences deposited into NCBI (Johnson et al. 2008).

The sequence query used in this investigation was obtained from the flowering plant *Leptosiphon jepsonii* (see Appendix B). This unique species exhibits flower-age-dependent self-incompatibility. The proportion of *Leptosiphon jepsonii* flowers that produce selfed progeny significantly increases as the flowers are allowed to mature (Goodwillie et al. 2004). RNAseq was performed on day one and day three flowers of *L. jepsonii* under the assumption that highly expressed genes present before the flowers mature, but present in smaller quantities after the flowers had matured, could be the genes responsible for self-incompatibility. To examine this, RNAseq was used to enrich for mRNAs over other RNAs. However, rRNAs are prevalent in all cells and are virtually always present in the output of RNAseq. The unwanted rRNA sequences

must be computationally removed before analyzing the batch of remaining sequences. However, the rRNAs were found to be abundant sequences in the *L. jepsonii* transcriptome, but matched closely with inferred protein sequences in automated BLASTX searches of NCBI. This phenomenon was the subject of investigation in my project.

Methods

After computationally identifying the most highly expressed genes in the transcriptome of *L. jepsonii,* we then analyzed the most highly expressed gene sequence using NCBI BLAST to determine whether a gene with a similar sequence or function had previously been identified and submitted to the database.

Using the BLAST program, the most abundant sequence obtained from *L. jepsonii* (see Appendix B) was input as the query and a BLASTX search was performed. BLASTX translates and compares a nucleotide query to a specified protein database. The BLASTX search provides a list of protein results (see Appendix A.II, A.III, and A.IV) that have statistically significant similarity to the nucleotide sequence used in the search, presumably derived from an mRNA. The most similar and least similar hits, with a cut-off e-value of e^{-10}, for a diverse list of eukaryotes (see Appendix C) were obtained and their accession numbers recorded. These accession numbers serve as a unique identifier for each sequence within the database. The e-value, or expect value, provides a significance threshold where the closer the value is to zero, the less probable it is that inferred alignment result in the BLAST search occurred by chance. The number of different organisms and the number of inferred protein accessions were also recorded from the hits associated with the BLASTX search (see Figure 2).

From each inferred protein recovered, reference sequence information could be used to obtain the mRNA accession number associated with the protein. A BLASTN search was run using this mRNA accession number as the search query. BLASTN is used to analyze a

nucleotide sequence against other nucleotide sequences, with no translation performed by the program (see Appendix A.I). This permitted matches to any type of RNA or DNA sequence in NCBI, not just those annotated as mRNAs encoding proteins. Rather than investigate every individual sequence recovered, the most similar and least similar search result was taken as a sampling strategy under the assumption that all the sequences above the e-value cutoff were likely rRNAs. That is, if both the top and bottom scoring sequences on the list both matched rRNA sequences, it is likely that most, if not all, of the sequences between these two are also rRNAs.

Lastly a BLASTP search was performed using the original protein hits as the search query and the model organisms *Homo sapiens*, *Drosophila melanogaster*, and *Arabidopsis thaliana* as the search set. BLASTP compares protein sequences by performing general sequence identification and similarity searches. Examining BLAST results in model organisms allowed for examination of accuracy of annotations within the most well characterized genomes.

Results and Discussion

The distribution of the BLASTX search results (Figure 2) indicates that, when searching the Viridiplantae set, there are more protein accession hits similar to the initial query sequence than when searching any other target group of organisms. This is expected because the query sequence was obtained from the organism *L. jepsonii*, a green plant itself. The search that yielded the second most protein accessions and organisms was that targeting Fungi. Examining the Metazoans also produced many protein accession and organism matches. Searching the Amoebozoa, Apicomplexa, Rhodophyta, and Stramenopile data sets produced some but substantially fewer results. The Kinetoplastida and Ciliata database searches returned no results (see Figure 2). Many of the results obtained exhibited a trend: they were annotated as "hypothetical protein," "uncharacterized protein," "unnamed protein," or "predicted protein" (see

Appendix A.III). It is unusual that the database categorizes these sequences as proteins sequences that are inferred across numerous organisms, including well-examined species, but have yet to be assigned any putative function.

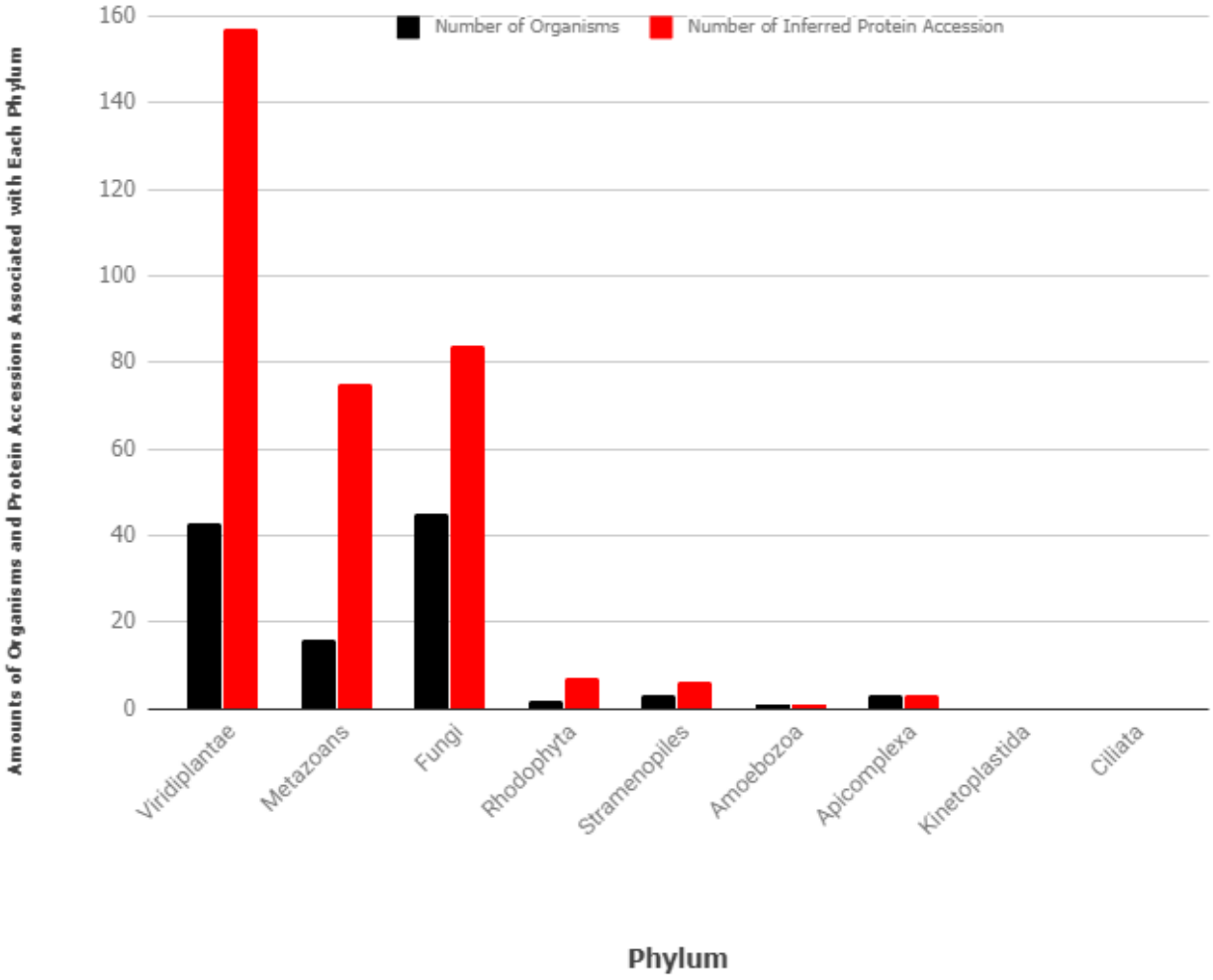


Figure 2. The number of individual organisms and the number of inferred protein accession hits for each eukaryotic phylum investigated within the statistically relevant cutoff range of greater than e^{-10}.

The BLASTN search query consisted of the mRNA RefSeq accession number associated with the protein hit from the BLASTX search (see Appendix A.V and A.VI). The most similar sequences retrieved by this search were expected to be other mRNA sequences because BLASTN compares a nucleotide sequence query with other nucleotide sequences within the database. However, rather than obtaining the expected mRNA results, the majority of the

statistically significant hits within an e-value cutoff of e^{-10} were rRNA sequences. There were few mRNA sequences hit, which could be rRNAs misannotated as mRNAs. If this is the case, misannotation may be present in all BLAST hits, not just the majority. These findings demonstrate the widespread misannotation of rRNA sequences spanning almost all major phyla of eukaryotes.

Using the model organisms *Homo sapiens*, *Arabidopsis*, and *Drosophila melanogaster*, the BLAST results were compared to determine the level of accuracy of annotations within the most well curated complete genomes. For the BLASTP search, the protein top and bottom hits from the original BLASTX search were used as the search query against the genomes of the model organisms. Most of these searches produced no results, or results with e-values outside the cutoff of e^{-10}, which is to be expected. However, some similarity existed between the BLASTX hits and protein sequence information from both *Homo sapiens* and *Arabidopsis*. For example, a protein found to be similar to the original search query using BLASTX was run through BLASTP with the model organisms as the target search organisms. This unexpected similarity indicates that rRNA sequences are misannotated as protein encoding genes even in the most well annotated genomes.

One limitation of this study is that of using only a plant sequence as the initial query. Had the search originated with an Animalia rRNA sequence missannotated as an mRNA, it is reasonable to expect more animal and fungi sequences would have been recovered. If a rhodophyte sequence was used as the search query, more sequences from rhodophytes would be expected. If animal sequences were searched with the less conserved ITS regions from animals, it is likely to find matches from animals within NCBI, but not matches to ITS regions from plants, fungi, or other more distantly related organisms. Part of the reason for the elevated level

of plant results is the rapid evolution of ITS regions compared to the slower evolution of the core rRNAs that are functional in the ribosome. Consequently, matches were found to plant ITS sequences missannotated as mRNAs, but other organisms are too distantly related for there to be enough similarity in ITS regions for BLAST searches to find. In other words, the results obtained using solely the *L. jepsonii* sequence very likely underestimate the severity of the misannotation problem.

Conclusion

The unexpected results obtained from the BLAST searches in this study indicate a quality control issue within the NCBI database, which very likely pervades other public databases. In recent years, the focus of modern research has been shifted from experimentally reproducible and rigorous results to "discovery-based" novel and exciting findings from genome-level sequencing. Combined with continuously increased speed and decreased cost of genome sequencing, due to the genomic revolution and the rapid technology developed thereafter, this shift has resulted in a massive "dumping" of sequences into public databases without proper validation of sequence annotations.

A large fraction of RNA recovered in metatranscriptomics is typically rRNA. It is well understood that the biosynthesis of ribosomes is a major cellular activity, particularly in association with cell growth (Koski & Golding 2001). For reliable results and meaningful future research, it is imperative that rRNA sequences, among other information available in public databases, be correctly annotated. Genomics researchers should reciprocally BLAST their results and double-check for validity before depositing incorrectly annotated sequences into the database. As a countermeasure, NCBI could consider implementing one or more novel quality control checkpoints in the BLAST program to ensure that rRNA sequences are not mistaken for

protein sequences (Tripp et al. 2011). Accurate annotations within metatranscriptomic databases like NCBI are essential for the scientific community to make reliable research progress. Basing new research studies on incorrect data from previous research leads to a cascade of spurious findings.

Until scientists “dumping” data into public databases, particularly but not necessarily limited to NCBI, address and resolve problematic widespread misannotation of rRNA sequences, researchers should be aware that thousands of misannotated rRNA sequences are easily revealed by a single reciprocal BLAST and will undoubtedly be present until procedures for curating and annotating data are reformed.

Acknowledgements

I thank Dr. John Stiller for his invaluable support and insight over the course of this project. Thank you to Carol Goodwillie for your research contribution that initiated my research. This study was made possible by the East Carolina University Department of Biology and the East Carolina University Honors College.

References

Goodwillie, C., Partis, K. L., & West, J. W. (2004). Transient Self-Incompatibility Confers Delayed Selfing in *Leptosiphon jepsonii*(Polemoniaceae). *International Journal of Plant Sciences*, 165(3), 387-394. doi:10.1086/382805

Jenkins, J., Grady, P. A., & Collins, F. S. (2005). Nurses and the Genomic Revolution. *Journal of Nursing Scholarship*, 37(2), 98-101. doi:10.1111/j.1547-5069.2005.00020.x

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., Mcginnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36(2), 5-9. doi:10.1093/nar/gkn201

Koski, L. B., & Golding, G. B. (2001). The Closest BLAST Hit Is Often Not the Nearest Neighbor. *Journal of Molecular Evolution*, 52(6), 540-542. doi:10.1007/s002390010184

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), 133-141. doi:10.1016/j.tig.2007.12.007

Ribosomes, Transcription, and Translation. Retrieved November 29, 2017, from https://www.nature.com/scitable/topicpage/ribosomes-transcription-and-translation-14120660

Transcriptome. Retrieved November 29, 2017, from https://www.genome.gov/13014330/transcriptome-fact-sheet/

Tripp, H. J., Hewson, I., Boyarsky, S., Stuart, J. M., & Zehr, J. P. (2011). Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Research*, 39(20), 8792-8802. doi:10.1093/nar/gkr576

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10, 57-63. doi:10.1038/nrg2484

Winnebeck, E. C., Millar, C. D., & Warman, G. R. (2010). Why Does Insect RNA Look Degraded? *Journal of Insect Science*, 10(159), 1-7. doi:10.1673/031.010.14119

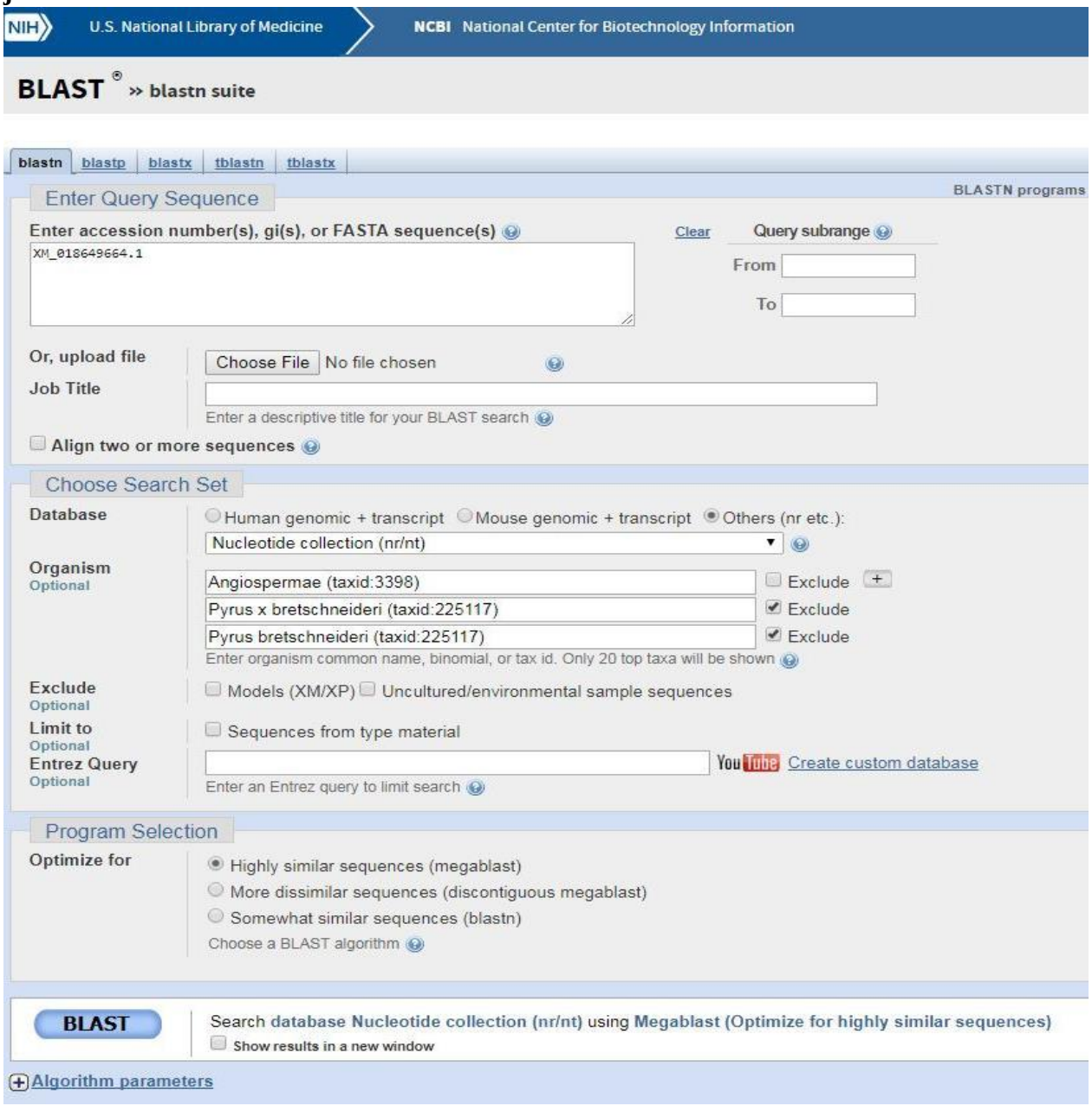
Appendix A

NCBI BLAST

The NCBI BLAST sequence similarity search program has several functions and various screenshots of the process used are shown below.

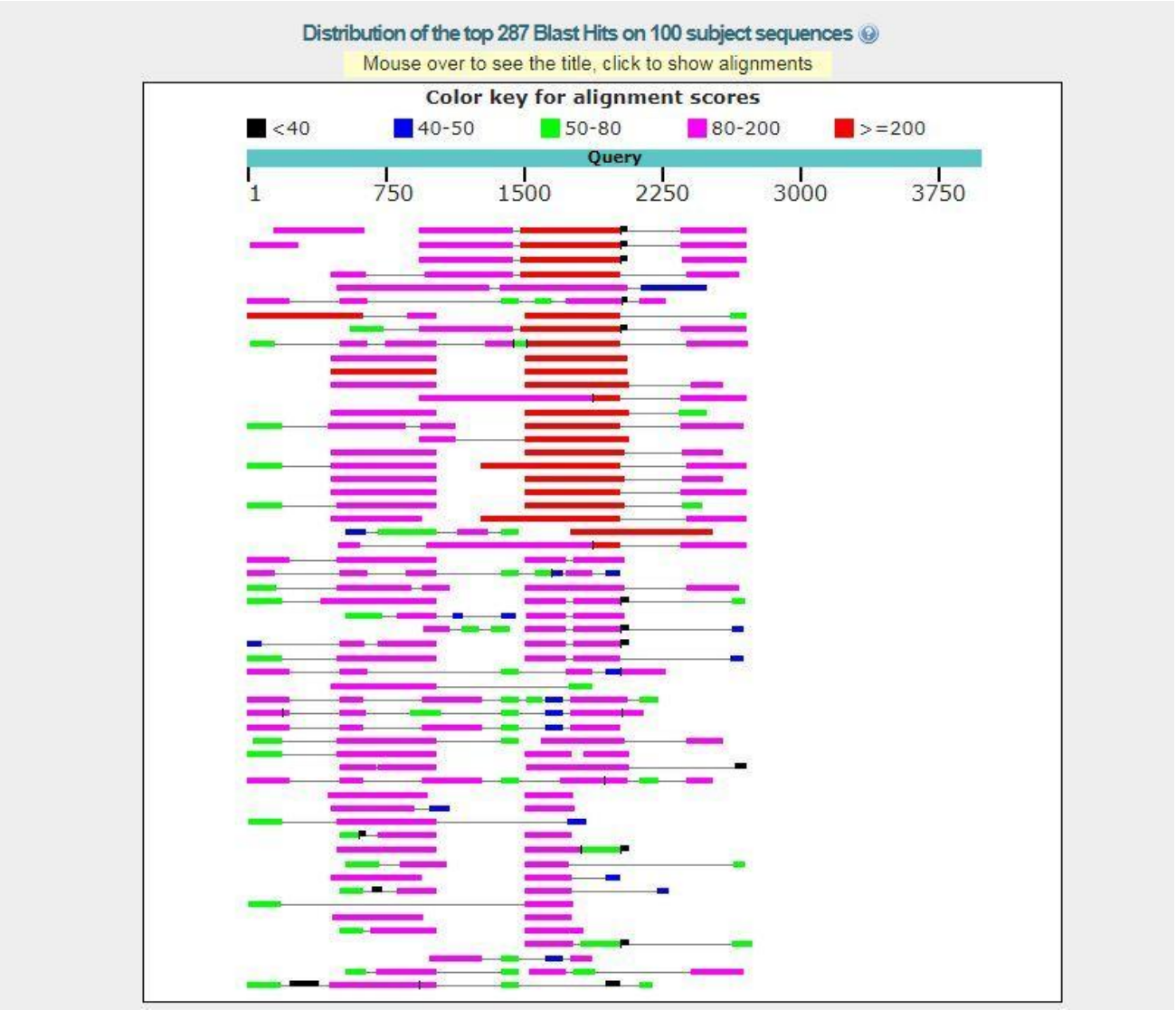
I.

john



NCBI BLAST search query

II.



NCBI BLAST results part 1- colorized visual representation of sequence alignments. The higher the alignment score, the closer the match to the query sequence.

III.

Sequences producing significant alignments:

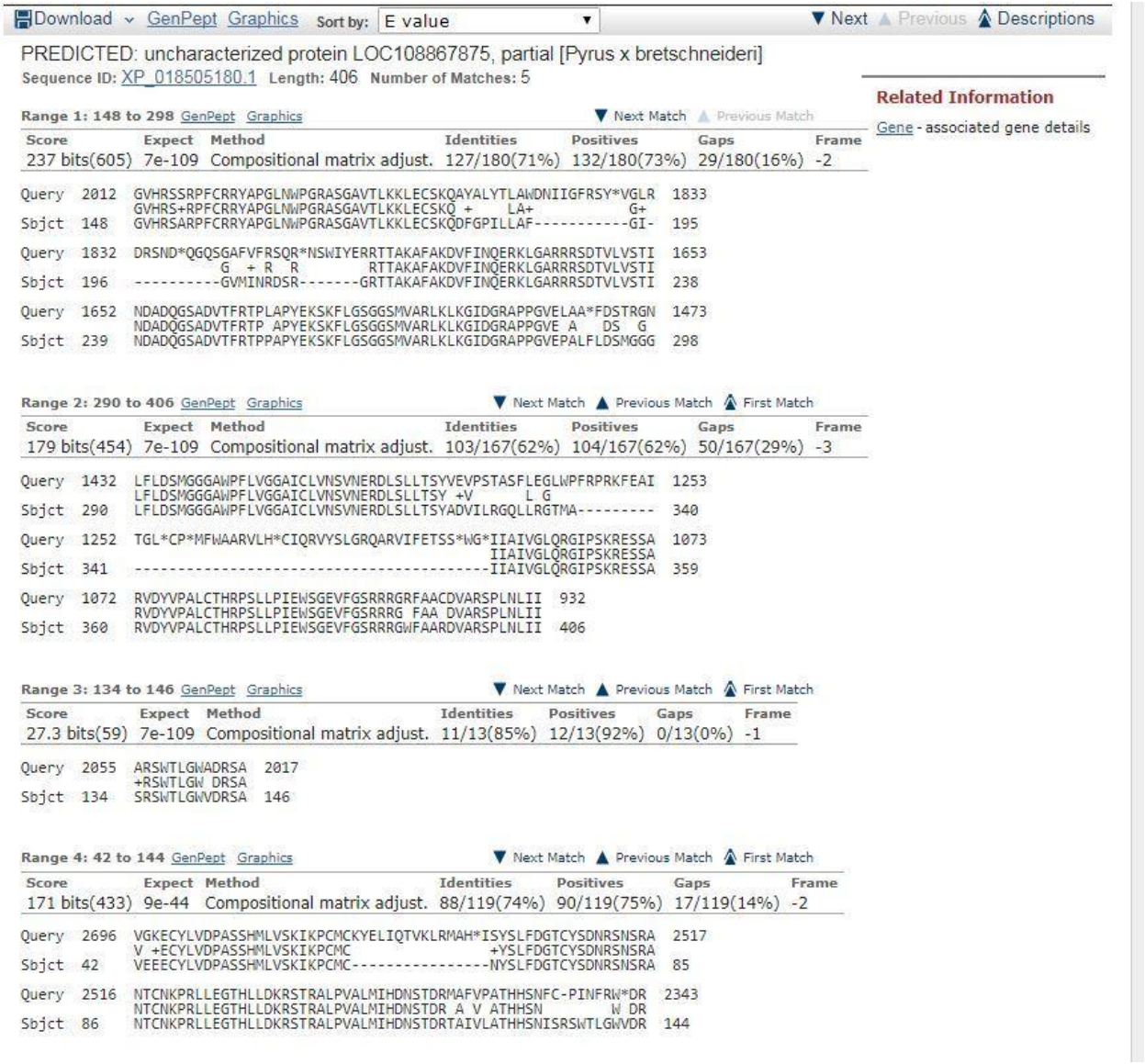
Select: All None Selected:0

Alignments Download GenPept Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: uncharacterized protein LOC108867875 [Pyrus x bretschneideri]	237	659	36%	7e-109	71%	XP_018505180.1
PREDICTED: uncharacterized protein LOC108867880 [Pyrus x bretschneideri]	234	655	36%	6e-108	70%	XP_018505183.1
PREDICTED: uncharacterized protein LOC108866852 [Pyrus x bretschneideri]	232	474	27%	3e-105	71%	XP_018501704.1
hypothetical protein GLYMA_13G012500 [Glycine max]	221	669	37%	2e-97	69%	KRH17750.1
uncharacterized protein LOC110926957 [Helianthus annuus]	186	392	47%	2e-93	52%	XP_022026266.1
hypothetical protein MTR_0055s0030 [Medicago truncatula]	142	561	26%	9e-81	73%	XP_013442961.1
hypothetical protein L484_002552 [Morus notabilis]	262	320	15%	3e-78	78%	XP_010106597.1
PREDICTED: uncharacterized protein LOC103956333 [Pyrus x bretschneideri]	231	878	40%	7e-78	71%	XP_009366572.1
hypothetical protein MTR_0055s0130 [Medicago truncatula]	226	833	40%	2e-74	69%	XP_013442970.1
uncharacterized protein LOC111289175 [Durio zibethinus]	228	228	14%	2e-66	66%	XP_022735763.1
uncharacterized protein LOC111291567 [Durio zibethinus]	226	226	14%	3e-66	65%	XP_022739110.1
PREDICTED: uncharacterized protein LOC108455178 [Gossypium arboreum]	225	337	18%	7e-65	64%	XP_017609266.1
PREDICTED: uncharacterized protein LOC108867878 [Pyrus x bretschneideri]	232	597	36%	7e-64	71%	XP_018505181.1
PREDICTED: uncharacterized protein LOC108455639 [Gossypium arboreum]	219	272	18%	9e-64	63%	XP_017609668.1
hypothetical protein GLYMA_13G016700 [Glycine max]	223	429	21%	2e-63	71%	KRH17795.1
PREDICTED: uncharacterized protein LOC105765883 [Gossypium raimondii]	218	347	19%	7e-62	64%	XP_012440607.1
PREDICTED: uncharacterized protein LOC108455646 [Gossypium arboreum]	216	367	19%	2e-61	64%	XP_017609674.1
uncharacterized protein LOC110275767 [Arachis duranensis]	222	473	27%	6e-61	55%	XP_020987670.1
PREDICTED: uncharacterized protein LOC108455641 [Gossypium arboreum]	214	365	19%	7e-61	62%	XP_017609669.1
uncharacterized protein LOC110277264 [Arachis duranensis]	216	419	22%	2e-60	68%	XP_020990109.1
PREDICTED: uncharacterized protein LOC105767074 [Gossypium raimondii]	204	271	16%	3e-58	61%	XP_012442053.1
uncharacterized protein LOC110277442 [Arachis duranensis]	211	462	27%	2e-57	54%	XP_020990217.1
hypothetical protein GLYMA_13G015500 [Glycine max]	201	293	19%	8e-57	52%	KRH17783.1
uncharacterized protein LOC110277292 [Arachis duranensis]	207	630	38%	6e-55	65%	XP_020990120.1
hypothetical protein MTR_0021s0160 [Medicago truncatula]	203	286	19%	1e-53	57%	XP_013443323.1
hypothetical protein CQW23_34927 [Capsicum baccatum]	199	258	14%	3e-53	63%	PHT25446.1
hypothetical protein BC332_33823 [Capsicum chinense]	200	200	14%	2e-52	63%	PHT97268.1
hypothetical protein CQW23_31306 [Capsicum baccatum]	88.6	274	17%	4e-52	78%	PHT29099.1
hypothetical protein EUTSA_v10028188mg [Eutrema salsugineum]	138	231	12%	1e-51	95%	XP_006405916.1

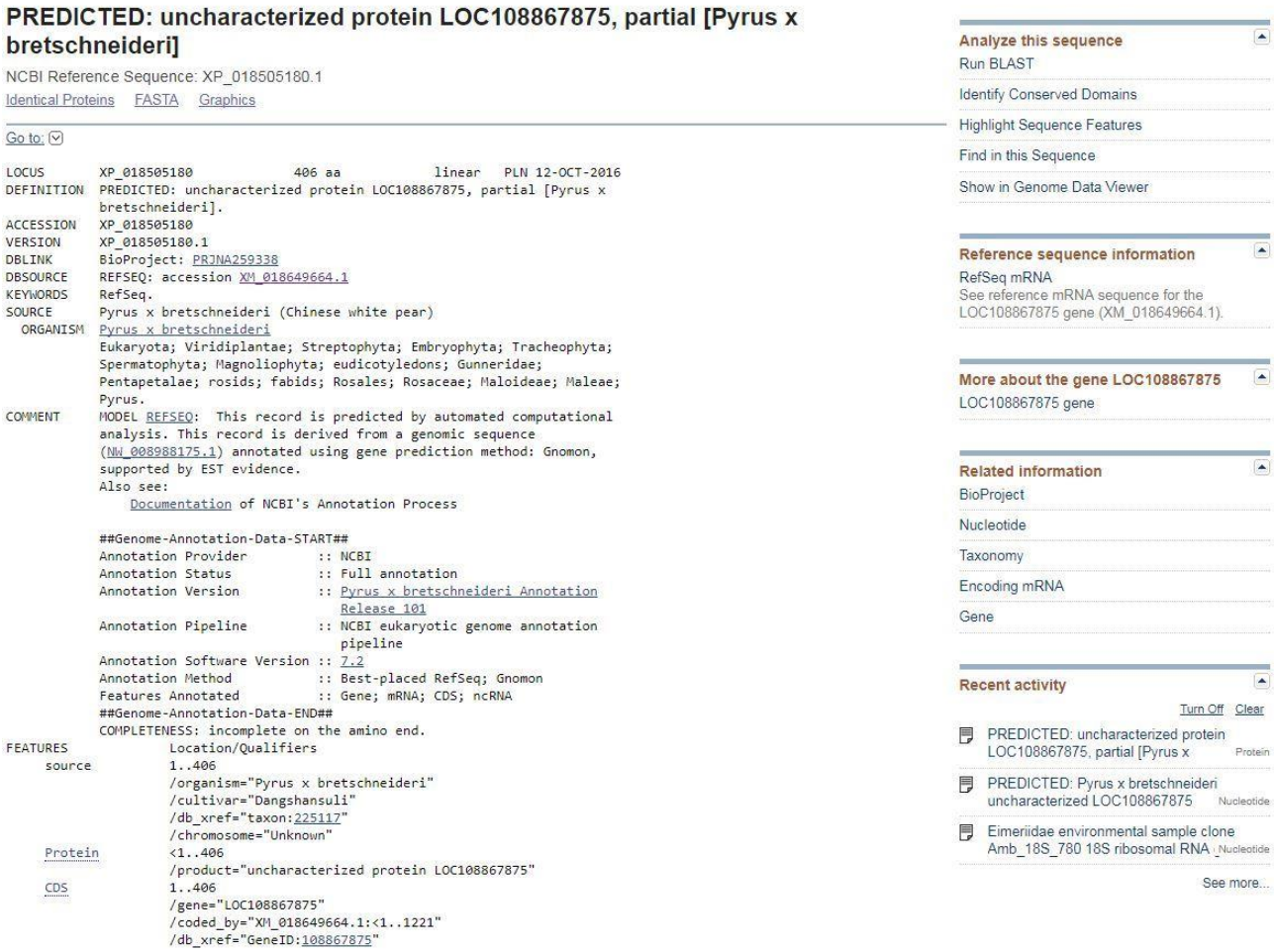
NCBI BLAST results part 2- list of top hits in order of e-value/similarity to the query sequence; note that almost all the results of this search are uncharacterized or hypothetical proteins.

IV.



NCBI BLAST results part 3- aligned sequence comparison between the query sequence and several rRNA sequences misannotated as proteins in NCBI.

V.



PREDICTED: uncharacterized protein LOC108867875, partial [Pyrus x bretschneideri]

NCBI Reference Sequence: XP_018505180.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       XP_018505180             406 aa            linear   PLN 12-OCT-2016
DEFINITION  PREDICTED: uncharacterized protein LOC108867875, partial [Pyrus x
            bretschneideri].
ACCESSION   XP_018505180
VERSION     XP_018505180.1
DBLINK      BioProject: PRJNA259338
DBSOURCE    REFSEQ: accession XM_018649664.1
KEYWORDS    RefSeq.
SOURCE      Pyrus x bretschneideri (Chinese white pear)
  ORGANISM  Pyrus x bretschneideri
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; fabids; Rosales; Rosaceae; Maloideae; Maleae;
            Pyrus.
COMMENT     MODEL REFSEQ:  This record is predicted by automated computational
            analysis. This record is derived from a genomic sequence
            (NW_008988175.1) annotated using gene prediction method: Gnomon,
            supported by EST evidence.
            Also see:
                Documentation of NCBI's Annotation Process

            ##Genome-Annotation-Data-START##
            Annotation Provider         :: NCBI
            Annotation Status           :: Full annotation
            Annotation Version          :: Pyrus x bretschneideri Annotation
                                           Release 101
            Annotation Pipeline         :: NCBI eukaryotic genome annotation
                                           pipeline
            Annotation Software Version :: 7.2
            Annotation Method           :: Best-placed RefSeq; Gnomon
            Features Annotated          :: Gene; mRNA; CDS; ncRNA
            ##Genome-Annotation-Data-END##
            COMPLETENESS: incomplete on the amino end.
FEATURES             Location/Qualifiers
     source          1..406
                     /organism="Pyrus x bretschneideri"
                     /cultivar="Dangshansuli"
                     /db_xref="taxon:225117"
                     /chromosome="Unknown"
     Protein         <1..406
                     /product="uncharacterized protein LOC108867875"
     CDS             1..406
                     /gene="LOC108867875"
                     /coded_by="XM_018649664.1:<1..1221"
                     /db_xref="GeneID:108867875"
```

Analyze this sequence
Run BLAST
Identify Conserved Domains
Highlight Sequence Features
Find in this Sequence
Show in Genome Data Viewer

Reference sequence information
RefSeq mRNA
See reference mRNA sequence for the LOC108867875 gene (XM_018649664.1).

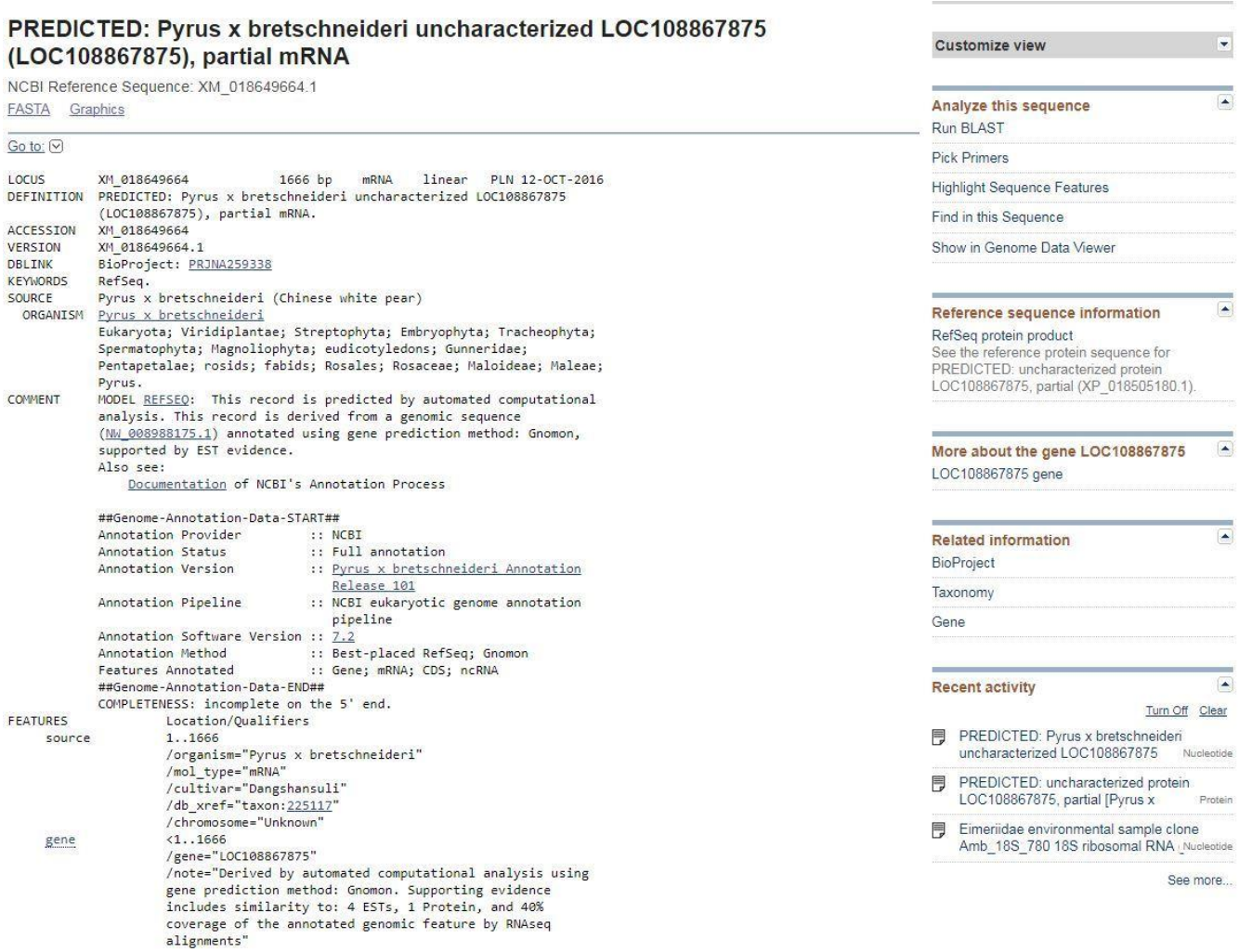
More about the gene LOC108867875
LOC108867875 gene

Related information
BioProject
Nucleotide
Taxonomy
Encoding mRNA
Gene

Recent activity
Turn Off Clear
PREDICTED: uncharacterized protein LOC108867875, partial [Pyrus x — Protein
PREDICTED: Pyrus x bretschneideri uncharacterized LOC108867875 — Nucleotide
Eimeriidae environmental sample clone Amb_18S_780 18S ribosomal RNA — Nucleotide
See more...

NCBI BLAST profile of protein result

VI.



PREDICTED: Pyrus x bretschneideri uncharacterized LOC108867875 (LOC108867875), partial mRNA

NCBI Reference Sequence: XM_018649664.1

FASTA Graphics

Go to:

```
LOCUS       XM_018649664            1666 bp    mRNA    linear   PLN 12-OCT-2016
DEFINITION  PREDICTED: Pyrus x bretschneideri uncharacterized LOC108867875
            (LOC108867875), partial mRNA.
ACCESSION   XM_018649664
VERSION     XM_018649664.1
DBLINK      BioProject: PRJNA259338
KEYWORDS    RefSeq.
SOURCE      Pyrus x bretschneideri (Chinese white pear)
  ORGANISM  Pyrus x bretschneideri
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; fabids; Rosales; Rosaceae; Maloideae; Maleae;
            Pyrus.
COMMENT     MODEL REFSEQ:  This record is predicted by automated computational
            analysis. This record is derived from a genomic sequence
            (NW_008988175.1) annotated using gene prediction method: Gnomon,
            supported by EST evidence.
            Also see:
                Documentation of NCBI's Annotation Process

            ##Genome-Annotation-Data-START##
            Annotation Provider         :: NCBI
            Annotation Status           :: Full annotation
            Annotation Version          :: Pyrus x bretschneideri Annotation
                                           Release 101
            Annotation Pipeline         :: NCBI eukaryotic genome annotation
                                           pipeline
            Annotation Software Version :: 7.2
            Annotation Method           :: Best-placed RefSeq; Gnomon
            Features Annotated          :: Gene; mRNA; CDS; ncRNA
            ##Genome-Annotation-Data-END##
            COMPLETENESS: incomplete on the 5' end.
FEATURES             Location/Qualifiers
     source          1..1666
                     /organism="Pyrus x bretschneideri"
                     /mol_type="mRNA"
                     /cultivar="Dangshansuli"
                     /db_xref="taxon:225117"
                     /chromosome="Unknown"
     gene            <1..1666
                     /gene="LOC108867875"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Gnomon. Supporting evidence
                     includes similarity to: 4 ESTs, 1 Protein, and 40%
                     coverage of the annotated genomic feature by RNAseq
                     alignments"
```

Customize view

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence
Show in Genome Data Viewer

Reference sequence information
RefSeq protein product
See the reference protein sequence for PREDICTED: uncharacterized protein LOC108867875, partial (XP_018505180.1).

More about the gene LOC108867875
LOC108867875 gene

Related information
BioProject
Taxonomy
Gene

Recent activity
Turn Off Clear
PREDICTED: Pyrus x bretschneideri uncharacterized LOC108867875 Nucleotide
PREDICTED: uncharacterized protein LOC108867875, partial [Pyrus x Protein
Eimeriidae environmental sample clone Amb_18S_780 18S ribosomal RNA Nucleotide
See more...

NCBI BLAST profile of Reference Sequence mRNA used to infer the annotated protein sequence.

Appendix B

Original Query Sequence Used

This sequence was obtained from the transcriptome of *Leptosiphon jepsonii*. This gene was the most highly expressed after the plant's flowers had reached sufficient maturity to be self-compatible.

>TRINITY_DN16035_c0_g1_i1
GCGCCTCGTGGTGCGACAGGGTCCGGACACGACGGGGCTCTCACCCTCTCTGGCGCCCCC
TTCCAGGGGACTTGGGTCCGGTCCGTCGCTGAGGACGCTTCTCCAGACTACAATTCGAAC
AACGAGGTCGTCCGATTTTCAAGCTGGGCTATTTCCGGTTCGCTCGCCGTTACTAAGGAA
ATCCTTGTAAGTTTCTTTTCCTCCGCTTATTGATATGCTTAAACTCAGCGGGTAGTCCCG
CCTGACCTGGGGTCGCGGTCGAGGCATCACTTTCGCAACGCAACAGGGTCATTGGAAACA
GCATAAGCAAGCGCACAACAGAGGCAAAGGTTTTAACAACCACCAATTGTTGTGACACTC
ACTCATGCCGACTCTTATTTAGGCCAACCACACAACATGTGTATGGGAGACCAATCTCCG
CTCCCAAAACCATGATTCCAAAAGGAATTGGTTGGGGGCGACGCGATGTGAGACGCCCAG
GCAGACGTGCCCTCGGCCTAATGGCTTCGGGCGCAACTTGCGTTCAAAAACTCGATGGTT
CACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAG
AGCCGAGATATCCGTTGCCGAGAGTCGTTTAGTTTTGTGTAAGATGGCATTGTCCTTTTG
CACCCGCGAACGAGGCTTCAAGACAAGCTCTCTTTGTTACAATTCCTTGGCACATTCCGT
GCCGGGGTTCGTTAAGTTGCCGGGAAGAACAACCATACACAAGGAATGGATGCTCCACCA
GCAAGGGGGCAACATACAAGTACCAAAGCACAAGTCATCACCCCCCAAGTTTGTGATACA
AGTTCGCGGGTCGTTCTGCTAGGCAGGTTTCGACAATGATCCTTCCGCAGGTTCACCTAC
GGAAACCTTGTTACGACTTCTCCTTCCTCTAAATGATAAGGTTCAGTGGACTTCTCGCGA
CGTCGCAGGCAGCGAACCGCCCACGTCGCCGCGATCCGAACACTTCACCGGACCATTCAA
TCGGTAGGAGCGACGGGCGGTGTGTACAAAGGGCAGGGACGTAGTCAACGCGAGCTGATG
ACTCGCGCTTACTAGGAATTCCTCGTTGAAGACCAACAATTGCAATGATCTATCCCCATC
ACGATGAAGTTTCAAAGATTACCCGGGCCTGTCGGCCAAGGCTATAAACTCGTTGAATAC
ATCAGTGTAGCACGCGTGCGGCCCAGAACATCTAAGGGCATCACAGACCTGTTATTGCCT
CAAACTTCCGTGGCCTGAAAGGCCATAGTCCCTCTAAGAAGCTAGCTGTGGAAGGGACTT
CCACATAGCTAGTTAGCAGGCTGAGGTCTCGTTCGTTAACGGAATTAACCAGACAAATCG
CTCCACCAACTAAGAACGGCCATGCACCACCACCCATAGAATCAAGAAAGAGCTCTCAGT
CTGTCAATCCTTACTATGTCTGGACCTGGTAAGTTTCCCCGTGTTGAGTCAAATTAAGCC
GCAAGCTCCACTCCTGGTGGTGCCCTTCCGTCAATTCCTTTAAGTTTCAGCCTTGCGACC
ATACTCCCCCCGGAACCCAAAAACTTTGATTTCTCATAAGGTGCCAGCGGAGTCCTAAAA
GTAACATCCGCTGATCCCTGGTCGGCATCGTTTATGGTTGAGACTAGGACGGTATCTGAT
CGTCTTCGAGCCCCCAACTTTCGTTCTTGATTAATGAAAACATCCTTGGCAAATGCTTTC
GCAGTTGTTCGTCTTTCGTAAATCCAAGAATTTCACCTCTGACTACGAAATACGAATGCC
CCCGACTGTCCCTGTTAATCATTACTCCGATCCCGAAGGCCAACTTAATAGGATCGAAAT
CCTATGATGTTATCCCATGCTAATGTATACAGAGCGTAGGCTTGCTTTGAGCACTCTAAT
TTCTTCAAAGTAACAGCGCCGGAGGCACGACCCGGCCAGTTAAGGCCAGGAGCGTATCGC
CGGCAGAAGGGACGAGACGACCGGTGCACACCGTAAGGCGGACCGGTCGGCCCAACCCAA
AGTCCAACTACGAGCTTTTTAACTGCAACAACTTAAATATACGCTATTGGAGCTGGAATT
ACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATGGATCCTCGTTAAGGGATTTAGATTG
TACTCATTCCAATTACCAGACTCAAAGAGCCCGGTATTGTTATTTATTGTCACTACCTCC
CCGTGTCAGGATTGGGTAATTTGCGCGCCTGCTGCCTTCCTTGGATGTGGTAGCCGTTTC
TCAGGCTCCCTCTCCGGAATCGAACCCTAATTCTCCGTCACCCGTCACCACCATGGTAGG
CCTCTATCCTACCATCGAAAGTTGATAGGGCAGAAATTTGAATGATGCGTCGCCGGCACA
AAGGCCATGCGATCCGTCGAGTTATCATGAATCATCAGAGCAACGGGCAAAGCCCGCGTC
GACCTTTTATCTAATAAATGCGTCCCTTCCAGAAGTCGGGGTTTGTTGCACGTATTAGCT
CTAGAATTACTACGGTTATCCGAGTAGCAGGTACCATCAAACAAACTATAACTGATTTAA
TGAGCCATTCGCAGTTTCACAGTCTGAATTAGTTCATACTTACACATGCATGGCTTAATC
TTTGAGACAAGCATATGACTACTGGCAGGATCAACCAGGTAGCATTCCTTTCCAACACGC

CAACAAACATGACTGTTCAACACTGAGATTGGTCAGCACATGCGTGAAGCGAGTCGTTCA
TGGTTTGCATCAAGGTAAAAGACGGTCATGGACCGCATACCCACAACATTTTCCGCATCC
TAAAGAACAAGCATCATCCCATGAGCCGTAGTCACAACACAATCAAGTGAAATAACACGG
AGCACAGGGTATGCCATTTGGTTCACCTCGCAGCAAAATGCAACAAGGCGAAATTCGCAA
CCGTTAAGAAGAAAAATTCCATCAATTTAGGTAGACAACACAGAAACCTAATTTGTTCCC
GTTGCATTTCAATACAATGGGATGACATTGAATAGGGTATGTTGGATATGGCTTGGCTGC
CGAGGCAGGGAACGACCAATCCAAACAAACCAAACACCACTCGTGCGTCATAACGTACAA
AATGGCATCAAATCCTCCCATCGAATCACACAGTCATGCAACCAAAATGGATGCACATCC
ATGCAAGAACGCCAAGAGAATCCAATGCCAGGTCCGATAGACACGTGAAACCAGGGACAA
GAAAGGCAGGCATCAAATCAAAGGCAACCACAACATTTTTGAACACAAGACAAGACATGC
AATTTTACATAACATTCATCACCATTATCCCTGAACATGGAAACAAGATGTGATTGGACA
ATGCCAATCAATCCATGCGTCACCAGAGGTCGTGCGGGTAACAGTAACTAATCGTATTTG
TTCAATTGCAAATCTAGTCCCAAGTCCTACCGTCACTGACTCAACAAGTTCCATCTTTTC
CACAATCTCATGCATTCAACCTTGAAAGGGTACGAAACCGCCAATACCAATCGCAAGCTA
ATGCACTTGAAGGAGGAAAATCCGTCACCCATCTAGTCCACACCCGAAATACAATTGAAA
GATTGTCCCTACGCCTGGCCTCCATGATTCCATAAATGGTGAAGGATCAACAGGTTTCTC
CGGGTAACGATAAATACCGTAGATCGATCACTTGTATTTTTTAGTTCAGGTGTTGGGACG
TTGTTGATCCCATTAGTCGTGGCCTCTAAAGAGGAGAGGGGCCACTACGTCTGGGATCCT
TGATGCCCACCATTGATGTTTTTCAACGAACTAGAAGCCATAGGCTTCCATGGCTTCAAA
AGCCATCGAAGCTATAGAAACCATAGGCTTCCAATGCCATGAAAAACCAAAGCCTTCCGT
CATTTGTTCACGAATGAACAAATTTCTTACTACAACCTGCTAGTAAGCTCTCCCCCCCCC
CCCCCG

Appendix C

Raw Data

The following is the raw data obtained and utilized for this research study. The image is a screenshot of the spreadsheet on which the data was compiled and organized.

Trinity Query	Search organism	Blastx	mRNA accession number	e value	Number of Organisms	Number of Inferred Protein Accession	Blastn	Type	e value	Blastp
TY_DN16035_c0	Viridiplantae	XP_018505180	XM_018649664	5.00E-109	43	157	KF800093	rRNA	0	no results
		XP_020705263	XM_020849604	1.00E-10			HM590391	rRNA	0	no results
	Metazoans	XP_022672560	XM_022816825 (honeybee mite)	1.00E-31	16	75	JN623081	rRNA	0	EAW73647 (Homo Sapiens)
		XP_001621448	XM_001621398 (starlet sea anemone)	1.00E-10			FJ913836	rRNA	4.00E-78	no results
	Fungi	XP_014564976	XM_014709490	5.00E-32	45	84	LK932024	rRNA	3.00E-73	XP_011523722
		XP_018289579	XM_018431832	1.00E-10			AF157159	rRNA	2.00E-98	no results
	Rhodophyta	XP_005708643	XM_005708586	8.00E-35	2	7	KJ907781	rRNA	9.00E-169	EAW73647
		XP_005711501	XM_005711444	3.00E-12			KF766115	rRNA	2.00E-46	NP_112558
	Stramenopiles	XP_009534205	XM_009535910	5.00E-19	3	6	AY742760	rRNA	0	no results
		XP_009534204	XM_009535909	2.00E-12			AB370108	rRNA	6.00E-138	no results
	Amoebozoa	XP_003295057	XM_003295009	1.00E-13	1	1	JN247435	rRNA	3.00E-54	no results
		N/A	N/A	N/A			N/A	N/A	N/A	N/A
	Apicomplexa	XP_012767848	XM_012912394	4.00E-26	3	3	HQ264136	rRNA	6.00E-75	NP_001330598 (Arabidopsis thaliana)
		XP_001610564	XM_001610514	4.00E-12			KC486858	rRNA	4.00E-55	BAC86923
	Kinetoplastida	no results	no results	no results	no results	no results	no results	no results	no results	no results
		no results	no results	no results	no results	no results	no results	no results	no results	no results
	Ciliata	1 result: e-value insignificant	no results	no results	no results	no results	no results	no results	no results	no results
		no results	no results	no results	no results	no results	no results	no results	no results	no results