GEP Annotation Report for D. Ananassae 1050L17

Anh-Dung Le

Student name:	Anh-Dung Le
Student email:	ale@unm.edu
Faculty advisor:	Dr. Paul Szauter
College/University:	University of New Mexico

Project Details

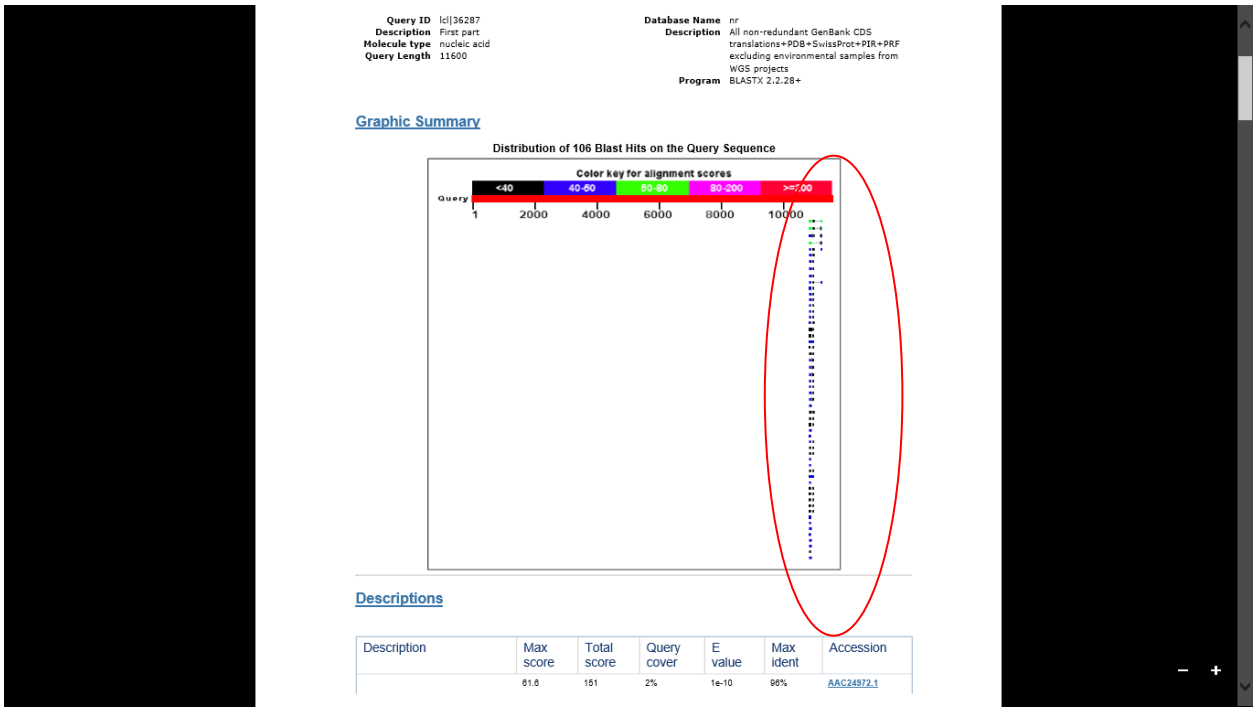
Project name:	dananassae_3Lcontrol_Jan2013_fosmid_1050L17
Project species:	*Drosophila ananassae*
Date of submission:	July 2, 2013
Size of project in base pairs:	38,247
Number of genes in project:	0

Project with no genes

1. BLASTX analysis

An unrestricted BLASTX scan was performed. The fosmid was isolated into 4 overlapping segments in order for the scan to be completed without crashing. The following are the results of the 4 scans:

part 1:



Query ID lcl|36287
Description First part
Molecule type nucleic acid
Query Length 11600

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTX 2.2.28+

Graphic Summary

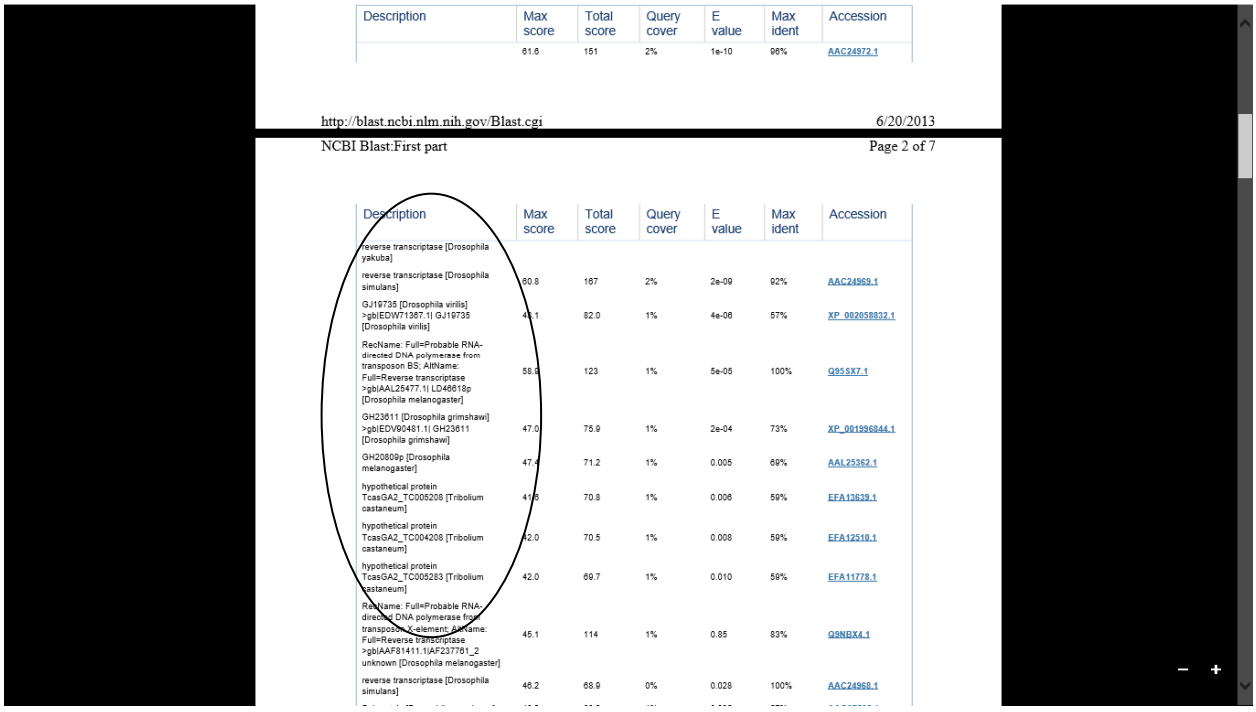
Distribution of 106 Blast Hits on the Query Sequence

Color key for alignment scores
<40 | 40-50 | 50-80 | 80-200 | >=200

Query
1 2000 4000 6000 8000 10000

Descriptions

Description	Max score	Total score	Query cover	E value	Max ident	Accession
	61.6	151	2%	1e-10	96%	AAC24972.1

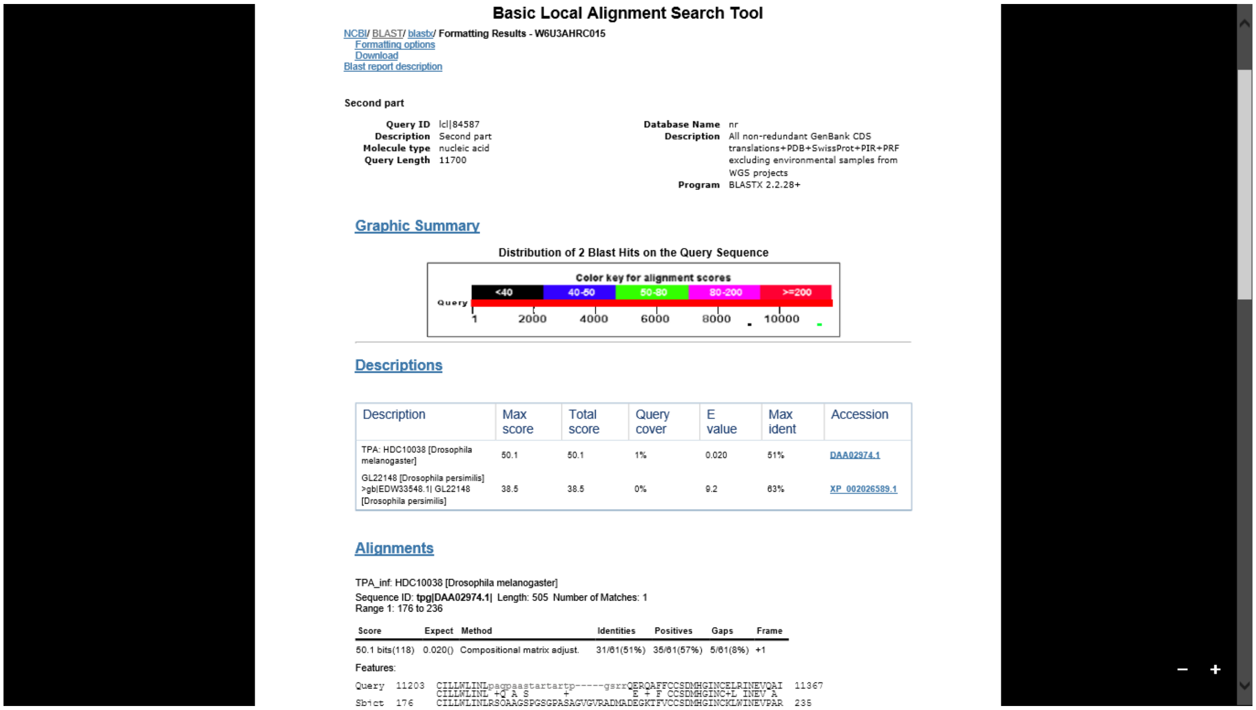


Description	Max score	Total score	Query cover	E value	Max ident	Accession
	61.6	151	2%	1e-10	96%	AAC24972.1

Description	Max score	Total score	Query cover	E value	Max ident	Accession
reverse transcriptase [Drosophila yakuba]						
reverse transcriptase [Drosophila simulans]	60.8	167	2%	2e-09	92%	AAC24969.1
GJ19735 [Drosophila virilis] >gb|EDW71367.1| GJ19735 [Drosophila virilis]	49.1	82.0	1%	4e-06	57%	XP_002058832.1
RecName: Full=Probable RNA-directed DNA polymerase from transposon BS; AltName: Full=Reverse transcriptase >gb|AAL25477.1| LD46618p [Drosophila melanogaster]	58.9	123	1%	5e-05	100%	Q95SX7.1
GH23611 [Drosophila grimshawi] >gb|EDV90481.1| GH23611 [Drosophila grimshawi]	47.0	75.9	1%	2e-04	73%	XP_001996844.1
GH20809p [Drosophila melanogaster]	47.4	71.2	1%	0.005	69%	AAL25362.1
hypothetical protein TcasGA2_TC005208 [Tribolium castaneum]	41.8	70.8	1%	0.006	59%	EFA13639.1
hypothetical protein TcasGA2_TC004208 [Tribolium castaneum]	42.0	70.5	1%	0.008	59%	EFA12510.1
hypothetical protein TcasGA2_TC005283 [Tribolium castaneum]	42.0	69.7	1%	0.010	59%	EFA11778.1
RecName: Full=Probable RNA-directed DNA polymerase from transposon X-element; AltName: Full=Reverse transcriptase >gb|AAF81411.1|AF237761_2 unknown [Drosophila melanogaster]	45.1	114	1%	0.85	83%	Q9NBX4.1
reverse transcriptase [Drosophila simulans]	46.2	68.9	0%	0.028	100%	AAC24968.1

Many of these genes point to reverse transcriptase or "RNA directed DNA polymerases". While many others are hypothetical genes.

part 2:



Basic Local Alignment Search Tool

NCBI/ BLAST/ blastx/ Formatting Results - W6U3AHRC015
Formatting options
Download
Blast report description

Second part

Query ID	lcl|84587	**Database Name**	nr
Description	Second part	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	nucleic acid		
Query Length	11700	**Program**	BLASTX 2.2.28+

Graphic Summary

Descriptions

Description	Max score	Total score	Query cover	E value	Max ident	Accession
TPA: HDC10038 [Drosophila melanogaster]	50.1	50.1	1%	0.020	51%	DAA02974.1
GL22148 [Drosophila persimilis] >gb|EDW33548.1| GL22148 [Drosophila persimilis]	38.5	38.5	0%	9.2	63%	XP_002026589.1

Alignments

TPA_inf: HDC10038 [Drosophila melanogaster]
Sequence ID: **tpg|DAA02974.1|** Length: 505 Number of Matches: 1
Range 1: 176 to 236

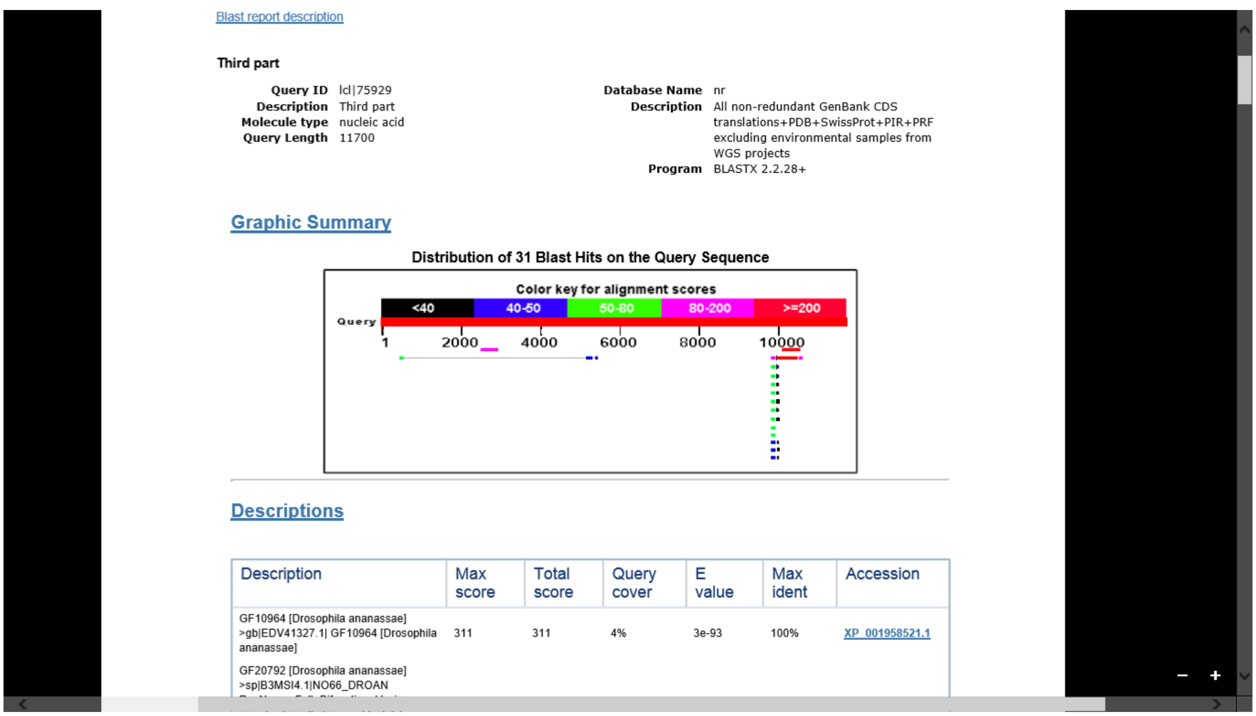
Score	Expect	Method	Identities	Positives	Gaps	Frame
50.1 bits(118)	0.020()	Compositional matrix adjust.	31/61(51%)	35/61(57%)	5/61(8%)	+1

Features:

```
Query  11203  CILLWLINLpagpaastartartp-----gsrrQERQAFFCCSDMHGINCELRINEVQAI  11367
              CILLWLINL +Q A S       +            E + F CCSDMHGINC+L INEV A
Sbjct  176    CILLWLINLRSQAAGSPGSGPASAGVGVRADMADEGKTFVCCSDMHGINCKLWINEVPAR  235
```

Only 2 short sequences aligned to part 2 of this fosmid. However, one sequence is a "third party annotation" and the other sequence is a model.

part 3:



Blast report description

Third part

Query ID	lcl|75929	**Database Name**	nr
Description	Third part	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	nucleic acid		
Query Length	11700	**Program**	BLASTX 2.2.28+

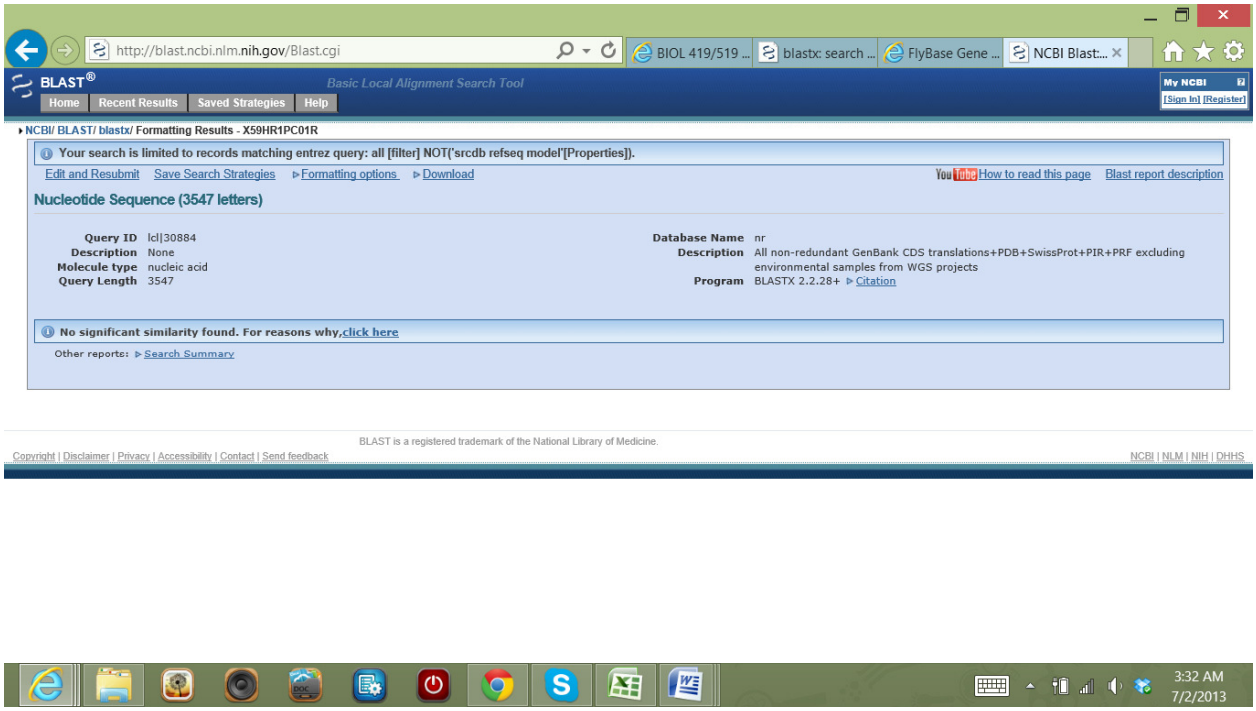
Graphic Summary

Descriptions

Description	Max score	Total score	Query cover	E value	Max ident	Accession
GF10964 [Drosophila ananassae] >gb|EDV41327.1| GF10964 [Drosophila ananassae]	311	311	4%	3e-93	100%	XP_001958521.1
GF20792 [Drosophila ananassae] >sp|B3MSI4.1|NO66_DROAN						

All of the alignments returned from the third section of the fosmid are either models or third party annotations except for one alignment. The returned alignment that is not a model nor a TPA was a gene labeled CG2982. There are several reasons why this not the gene which the fosmid contains. First, the aligned segment of only about 70 AA shows only a mediocre match. Second, Flybase indicates the gene is on the X chromosome which is inconsistent with D. Ananassae. Third, the gene was not predicted by GENSCAN. While this may not be the sole reason as to why this is not the gene, it is a certainly one of the many reasons. Fourth, based on an overall BLASTX of the fosmid, the gene is located near the 33 kb region which does not match up with the gene expression tracks indicated by Top Hat in UCSC Genome Browser.

Part 4:



Based on all four parts of the fosmid, no genes were matched by BLASTX with significant considerations.

2. Unrestricted BLASTP of GENSCAN predictions

Query	Top Hit		E	Coverage	Max identity
	Accession	Gene			
GENSCAN_predicted peptide_1	XP_003844655.1	predicted protein [Leptosphaeria maculans JN3]	1	22%	36%

GENSCAN_predicted peptide_2	YP_007311286.1	proteasome Rpn11 subunit JAMM motif protein	7.1	41%	29%
GENSCAN_predicted peptide_3	NP_001011115.1	serum response factor-binding protein 1 [Xenopus (Silurana) tropicalis]	5.00E-25	75%	51%
GENSCAN_predicted peptide_4	DAA02974.1	TPA: HDC10038 [Drosophila melanogaster]	8.00E-03	8%	88%
GENSCAN_predicted peptide_5	XP_002390526.1	hypothetical protein MPER_10178 [Moniliophthora perniciosa FA553]	3.8	30%	33%

3. Gene Expression Tracks Examination

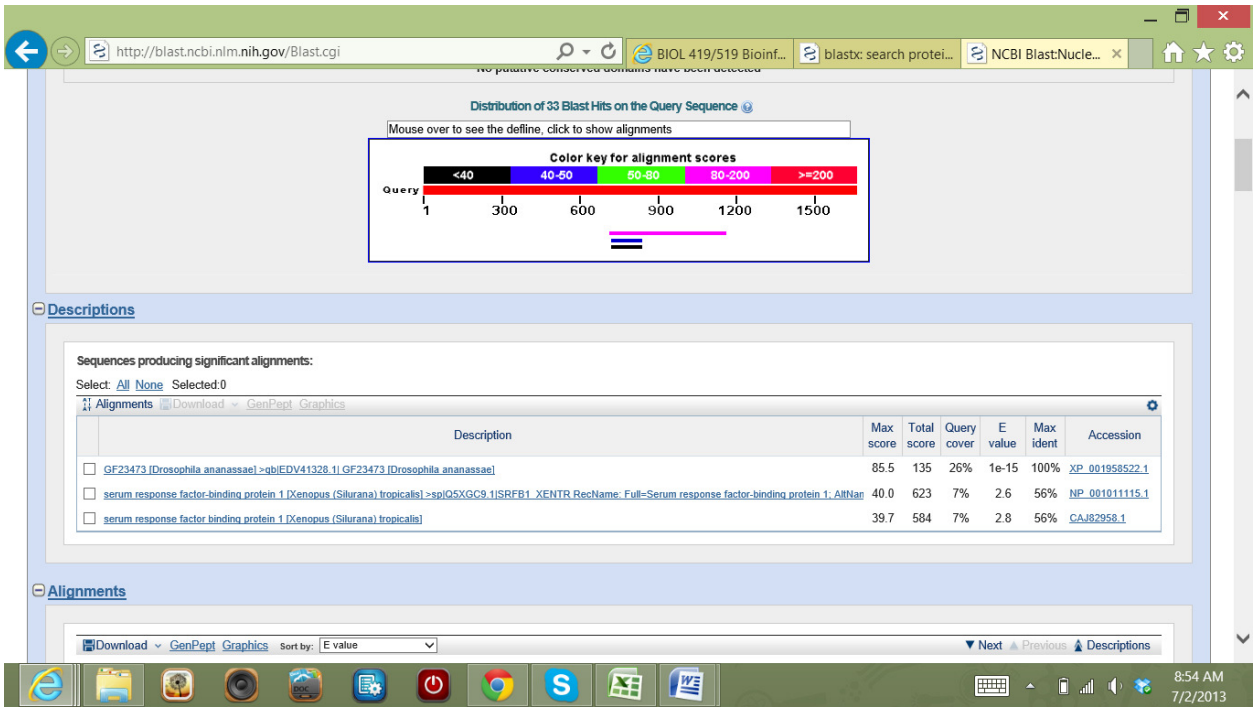
The gene expression tracks in UCSC Genome Browser shows evidence of transcription in the region of 26 kb.

TopHat track shows evidence of only transcripts from the minus strand. There is a lot apparent alternative splicing, however none of which are alignments to proteins. When we zoom in to this region, we can collect the transcribed region being generous about the ends:

>mystery_gene

TGGTATTTGATTGCTTAACACATATGTTCGTTAGAAGAATTGCAATTCTA
ATGTCTCAGGGAGATCATTTCTAAAGGTGTACACCTGAAGGACCTGGAGG
TTGAGAGAACAAAAAACGTGAATTCATAAATATATCACTGAAGAACCAGG
GGATTTGCAAAGATCGTCGTAAGTCAAGGGGCTCGGCTGTAGCCAATACA
ATTTTTGATTTGAAAAAGCTCCGATTATACTAGAAAGTTCCACTCTTTCA
ATTAACACGCACATTATGCAACGGCCAATAATCTATTTAGTGTTCTCGGG
TATATATAAGCGGAGCAATGCCGATTTATTTAGAGTATAGTTGGTCATAA
TGCAAAATCTTCTGCCTCTGGGCCCGCTCCCAGTTCCTTCTAGCTGCGGA
CTTAACGAAGTTTGGAATTCCTTAAGCTCCGCCTGCTTACCATTTCAATA
TAACGAGCGTGATGCGTTTTATCCTTACATGCAGGATCCTTATGCAGGGT
TCAGGTCCAAAGATGTATTAAGAGGCACTGGAGGCCCCGGTGGAAATGGA
GGCCCTGGAGGCCCTGGAGGACCTGGAGGACCAGGTGGCCCAGGAGGACC
TGGTGGGCCTGGAGGACCCGGAGGACCTGGAGGGGATGGAGGACATGGAG
GGCATGGAGGTCACGGAGGGCATGGAGGACCTGGAGGAGTTGGAGGACCT
GCCATTGTTGGAAATGGCGGTGGTGGAAATGTTGGCGGCGGAAAACCTGG
TGGTGGAAATGCCGGCGGTGGAAAACCTGGTGGTGGAAATGCCGGCGGTG
GAAAACCCAGTGGTGGAAATGTCGGCGATGGAAAACCTGGTAGTGGAAAT
GACGGAGGTGGAAATGTCGGTGATGCAAAACCTGGTGGAGGAAATGCCGG
CGGTGGAAATGCCGGTGGTGGAAATGCCGGCGGTGGAAAACCTGGTGGTG
GAAATGTCGGTGGTGGAAATTCCGGTGGTGGAAAACCTGGTGGTGGAAAT
GCCGGCGGTGGAAATGCCGGTGGTGGAAATGCAAGTGGTGGAAATGTCGG
CACTGGAAAACCTGGTGGTGGAAATGGCGGTGGTGGAAATGCCGGCGGTG
GAAAACCTGGTGGTGGAAATTCCGGCGGTGGAAAACCTGGTGGTGGAAAT
GCCGGCGGTGGAAAACCTGGCGTTGGAAGTTTAGGTGGAAGCATCGGCGG
TGGTCCATCTATAAAGCCTCCAATAGGACCAACAATAAGACCTCCTAAAG
GAGGTAATATAAAACCCCCCGTGGTAGGTGAAATAGAGACGCCGGATATA
CCTTACATTGAACCTCCTACCTGTCCCCCCATTGAGGTACCTATCTGGGA
TCCAAATTTGAATATATGGACCCAAAACAAAAGCCCTCGCTGTGTTTCCA
AAAATACAAAAAATTAAAAAAAAATTATATTCATTTGAAATGATTGTTTT
TTTAGTGTTACAGTTATAGTTATGGTTAGTTTAGACTAGTGTACCAATAG
AGAAAGTCCCAAAAACCCTTAAACCCCATAGATATGAATATCTCCTTCGA
CCCAGCCATTCTTAATTATATAGTCAAAAAAAGGTCTGATCTGAAGGCAG
ATCTGGAGTTGATGGCTTGTCAACTGAGGAACTATGTTGCGTAGAAACGG

A BLASTX vs. nr is performed using this sequence:



Only one significant match showing XP_001958522.1 which is a computational model from D. Ananasae.