GEP Annotation Report

Student name: Kaitlin Leddy
Student email: kleddy@unm.edu
Faculty Advisor: Dr. Paul Szauter
College/University: University of New Mexico

Project details

Project name: *D. ananassae* fosmid 2602K20
Project species: *D. ananassae*
Date of submission: 7-2-13
Size of project in base pairs: 41,165
Number Of genes in project: 2

Does this report cover all genes and all isoforms or is it a partial report? Full Report

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. ananassae* CG43392
Gene symbol (i.e.dmoj_ey): Dana\CG43392
Approximate location in project (from 5' end to 3' end): 33170-33406
Number of isoforms in *D. melanogaster*: 1
Number of isoforms in this project: 1

Complete the following table for all the isoforms in this project:

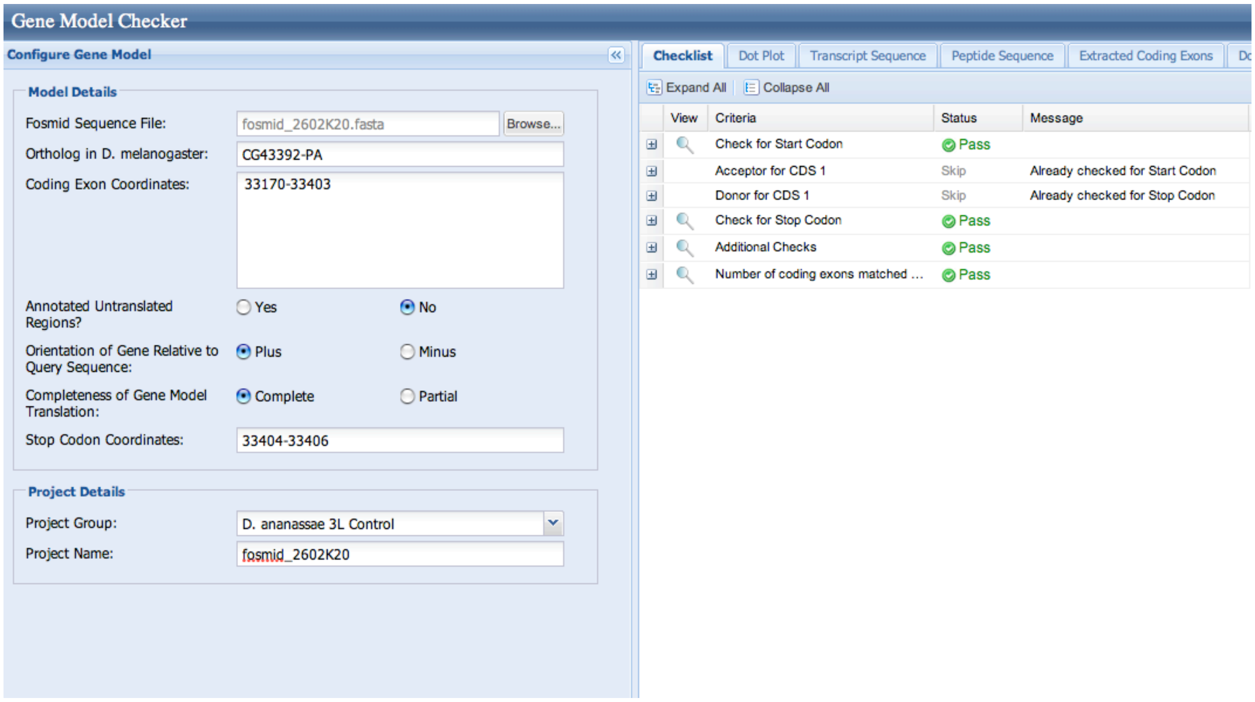
Name of unique isoform based on coding sequence	List of isoforms with identical coding sequences
Dana\CG43392-PA	**None**

Isoform report form

Gene--- isoform name (i.e. dmoj_ey--- PA): Dana\CG43392-PA
Names of the isoforms with identical coding sequences as this isoform
None

Is the 5' end of this isoform missing from the end of project: No
Is the 3' end of this isoform missing from the end of project: No

1. Gene Model Checker checklist



Model Details

Fosmid Sequence File: fosmid_2602K20.fasta

Ortholog in D. melanogaster: CG43392-PA

Coding Exon Coordinates: 33170-33403

Annotated Untranslated Regions? ○ Yes ● No

Orientation of Gene Relative to Query Sequence: ● Plus ○ Minus

Completeness of Gene Model Translation: ● Complete ○ Partial

Stop Codon Coordinates: 33404-33406

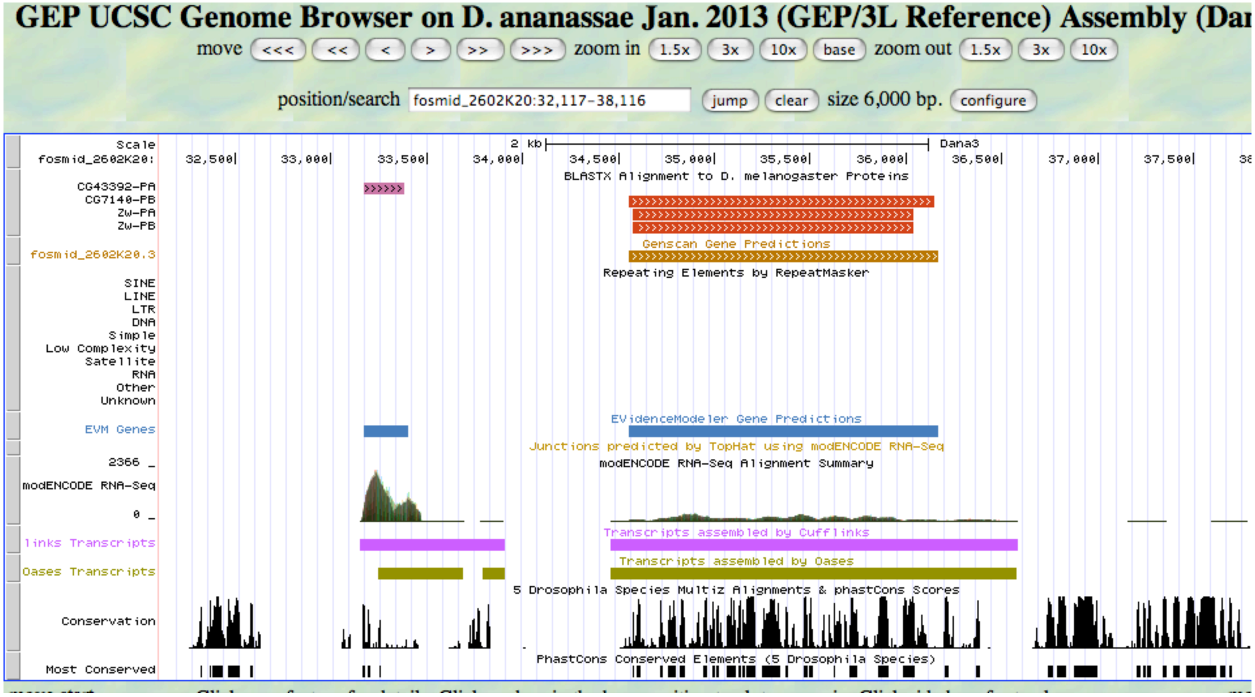
Project Details

Project Group: D. ananassae 3L Control

Project Name: fosmid_2602K20

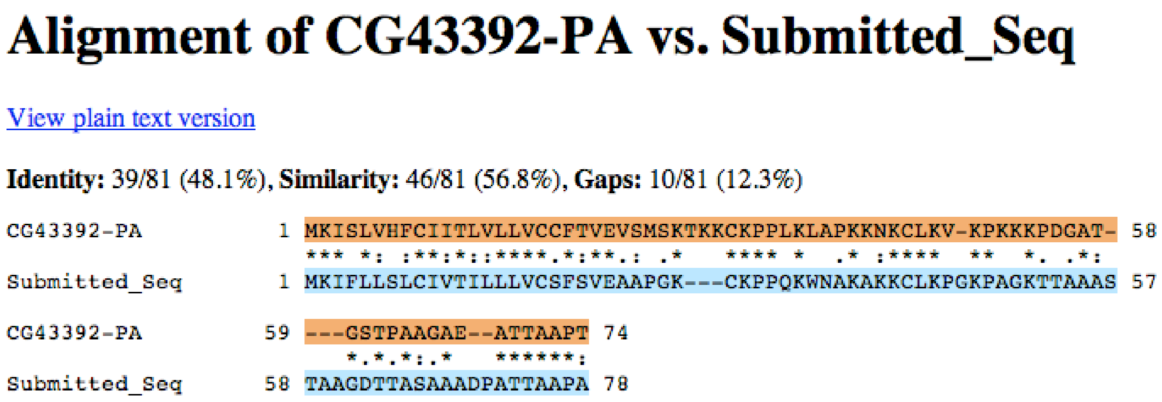
View	Criteria	Status	Message
🔍	Check for Start Codon	✓ Pass	
	Acceptor for CDS 1	Skip	Already checked for Start Codon
	Donor for CDS 1	Skip	Already checked for Stop Codon
🔍	Check for Stop Codon	✓ Pass	
🔍	Additional Checks	✓ Pass	
🔍	Number of coding exons matched ...	✓ Pass	

2. View the gene model on the Genome Browser



position/search fosmid_2602K20:32,117-38,116 size 6,000 bp.

3. Alignment between the submitted model and the *D. melanogaster ortholog*



Alignment of CG43392-PA vs. Submitted_Seq

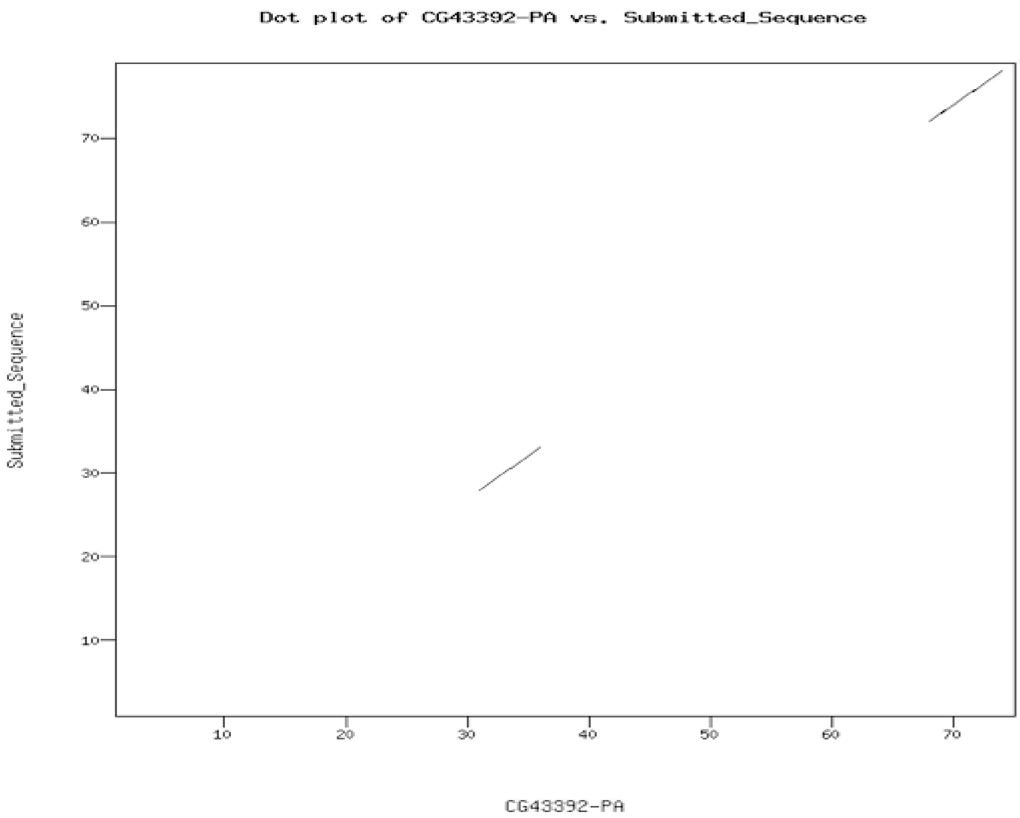
View plain text version

Identity: 39/81 (48.1%), **Similarity:** 46/81 (56.8%), **Gaps:** 10/81 (12.3%)

```
CG43392-PA            1 MKISLVHFCIITLVLLVCCFTVEVSMSKTKKCKPPLKLAPKKNKCLKV-KPKKKPDGAT- 58
                        *** *: :**:*::****.*:**.: .*   **** *  .* :****  **  *. .*:
Submitted_Seq         1 MKIFLLSLCIVTILLLVCSFSVEAAPGK---CKPPQKWNAKAKKCLKPGKPAGKTTAAAS 57

CG43392-PA           59 ---GSTPAAGAE--ATTAAPT 74
                           *.*.*:.*   ******:
Submitted_Seq        58 TAAGDTTASAAADPATTAAPA 78
```

---

4. Dot plot between the submitted model and the *D. melanogaster ortholog*



Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. ananassae* CG7140
Gene symbol (i.e.dmoj_ey): Dana\CG7140
Approximate location in project (from 5' end to 3' end): 34557-36167
Number of isoforms in *D. melanogaster*: 1
Number of isoforms in this project: 1

Complete the following table for all the isoforms in this project:

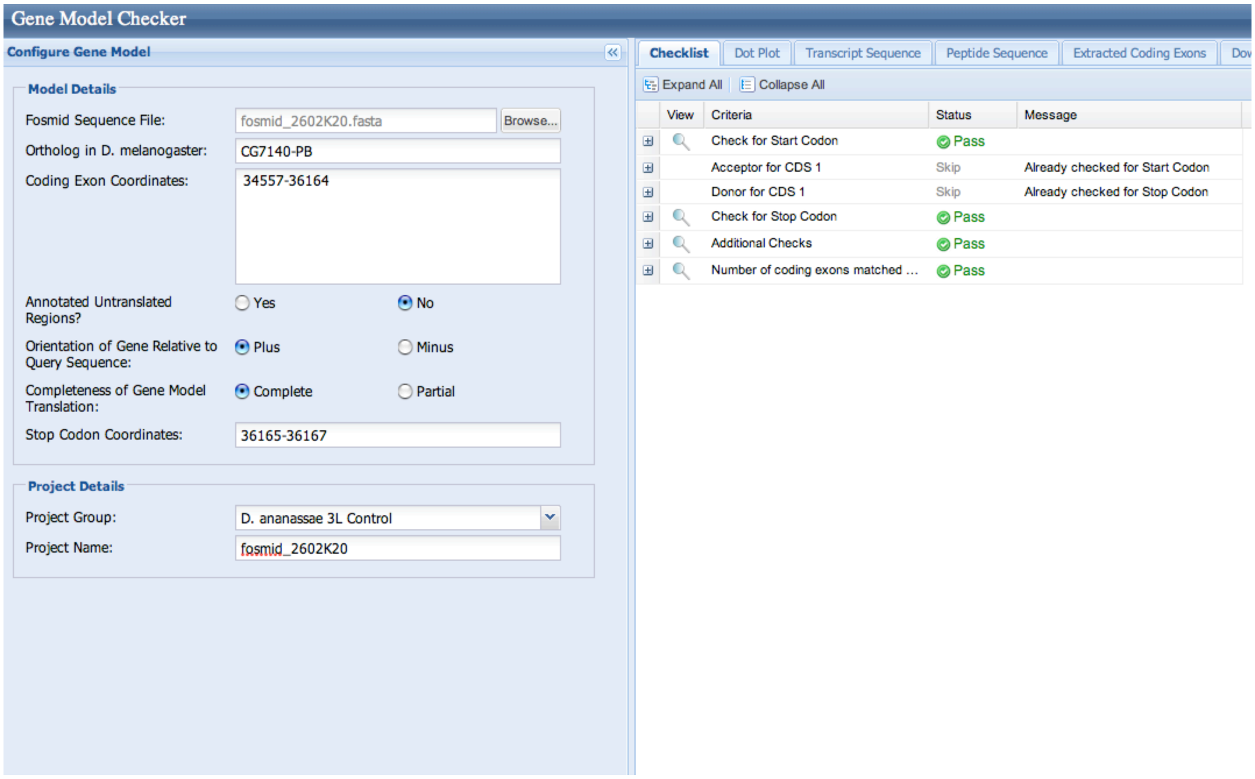
Name of unique isoform based on coding sequence	List of isoforms with identical coding sequences
Dana\CG7140	None

Isoform report form

Gene-‐‐ isoform name (i.e. dmoj_ey-‐‐ PA): Dana\CG7140
Names of the isoforms with identical coding sequences as this isoform
None

Is the 5' end of this isoform missing from the end of project: No
Is the 3' end of this isoform missing from the end of project: No

1. Gene Model Checker checklist



Model Details

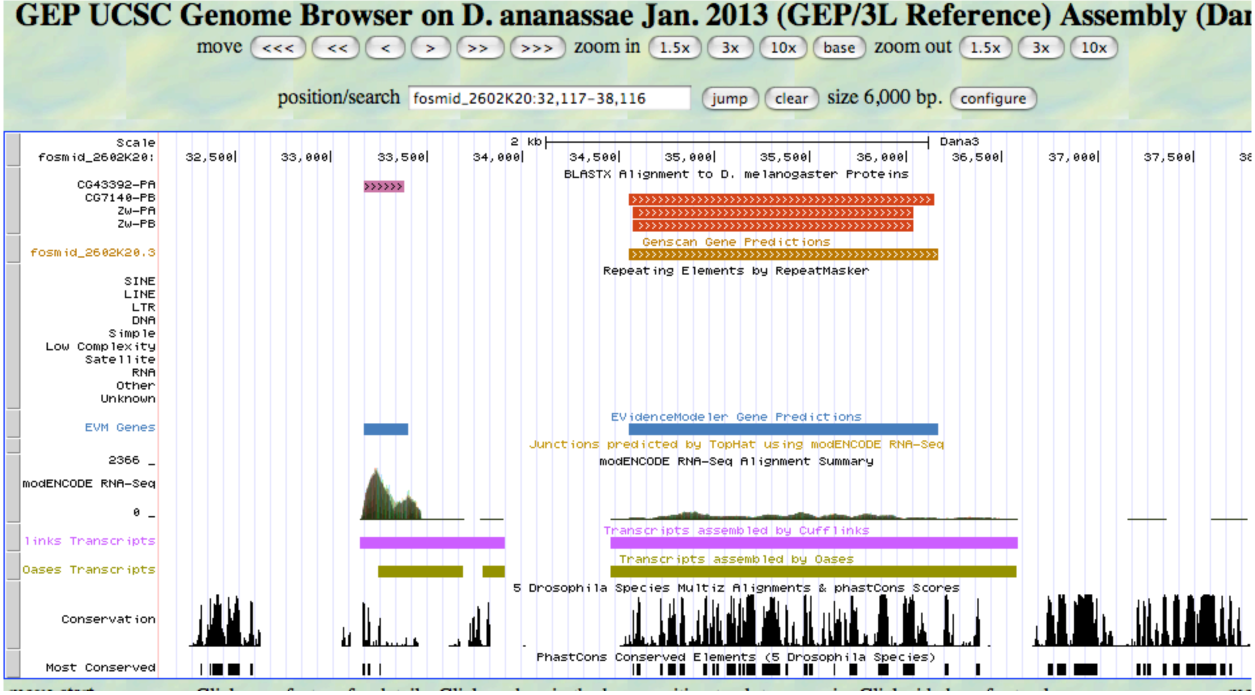
Fosmid Sequence File: fosmid_2602K20.fasta
Ortholog in D. melanogaster: CG7140-PB
Coding Exon Coordinates: 34557-36164
Annotated Untranslated Regions? ○ Yes ● No
Orientation of Gene Relative to Query Sequence: ● Plus ○ Minus
Completeness of Gene Model Translation: ● Complete ○ Partial
Stop Codon Coordinates: 36165-36167

Project Details

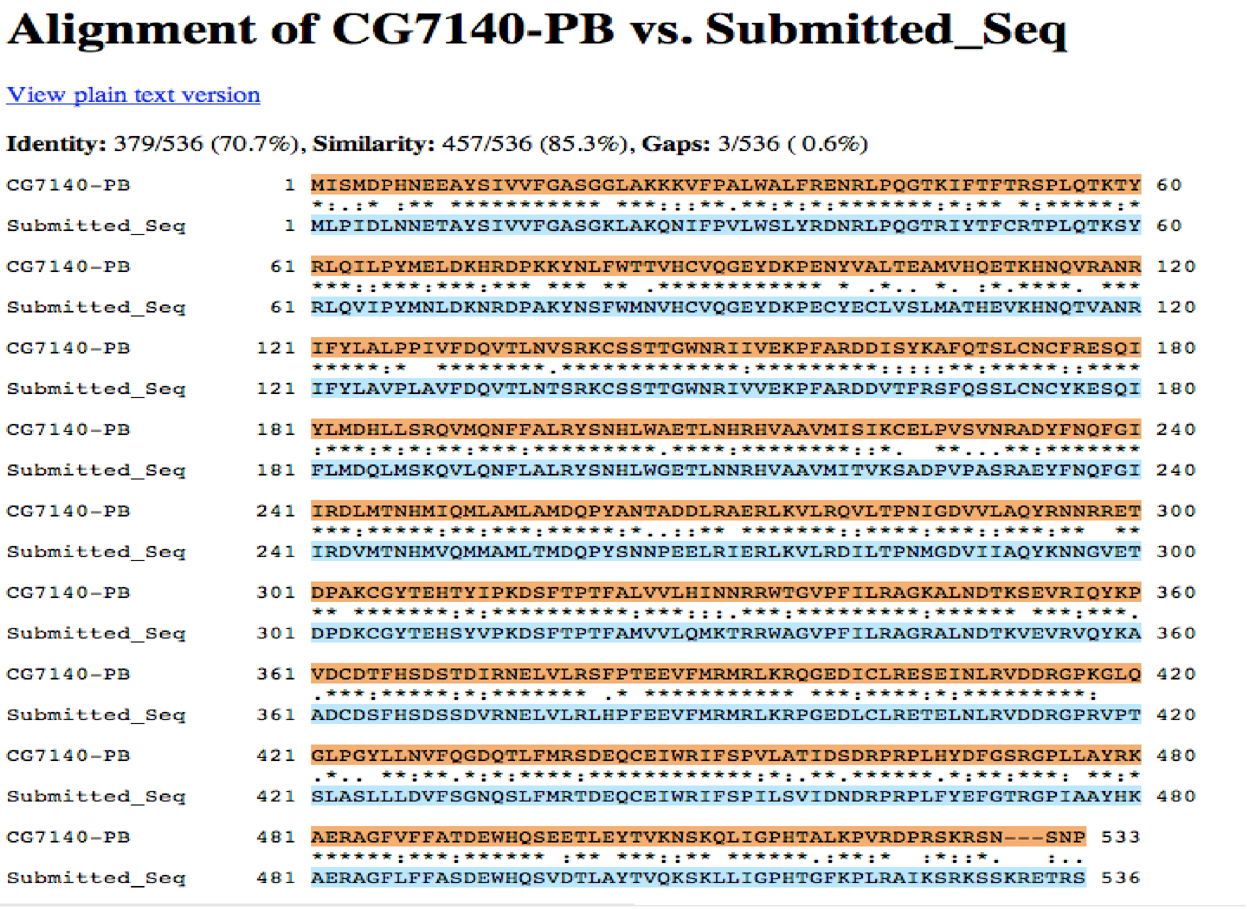
Project Group: D. ananassae 3L Control
Project Name: fosmid_2602K20

View	Criteria	Status	Message
🔍	Check for Start Codon	✓ Pass	
	Acceptor for CDS 1	Skip	Already checked for Start Codon
	Donor for CDS 1	Skip	Already checked for Stop Codon
🔍	Check for Stop Codon	✓ Pass	
🔍	Additional Checks	✓ Pass	
🔍	Number of coding exons matched ...	✓ Pass	

2. View the gene model on the Genome Browser



3. Alignment between the submitted model and the *D. melanogaster ortholog*



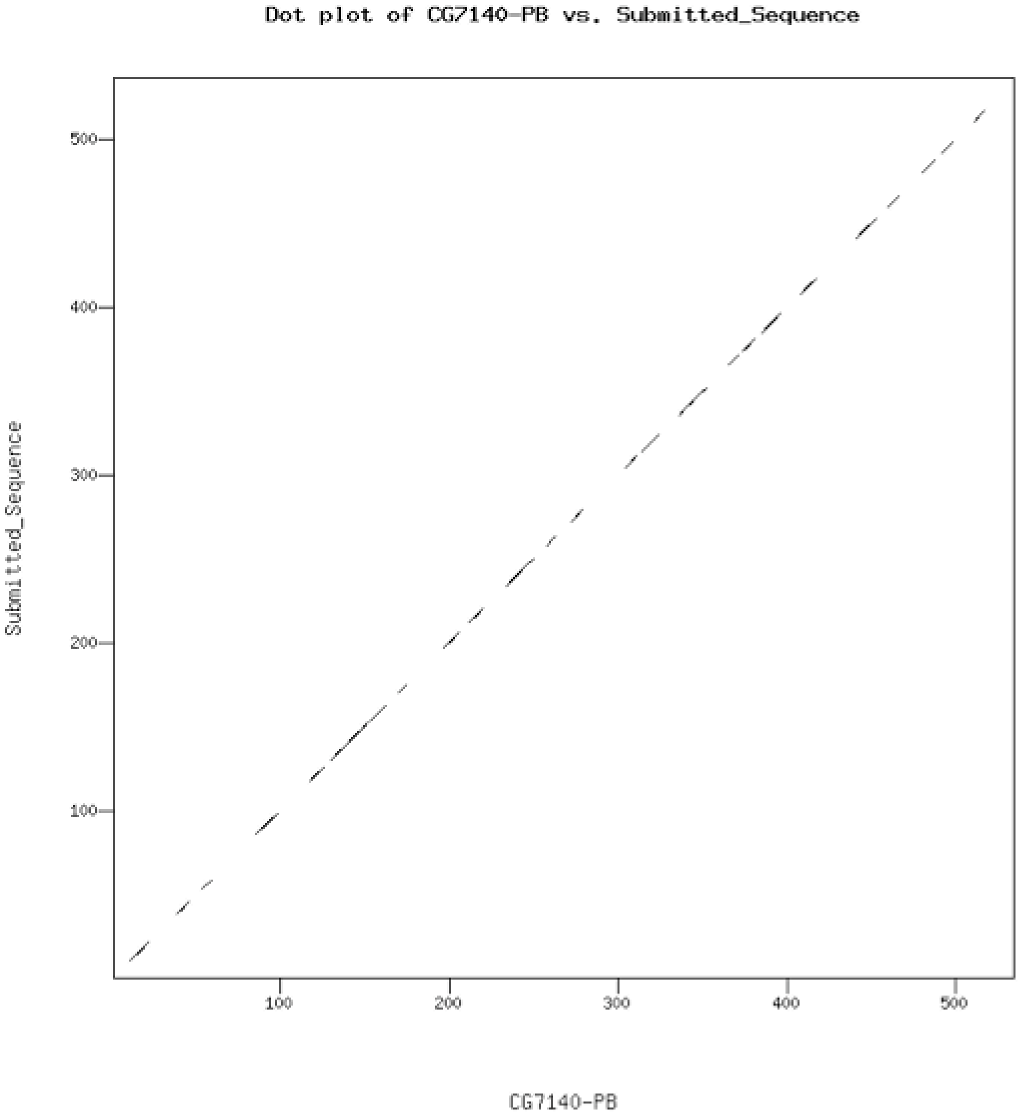
Alignment of CG7140-PB vs. Submitted_Seq

View plain text version

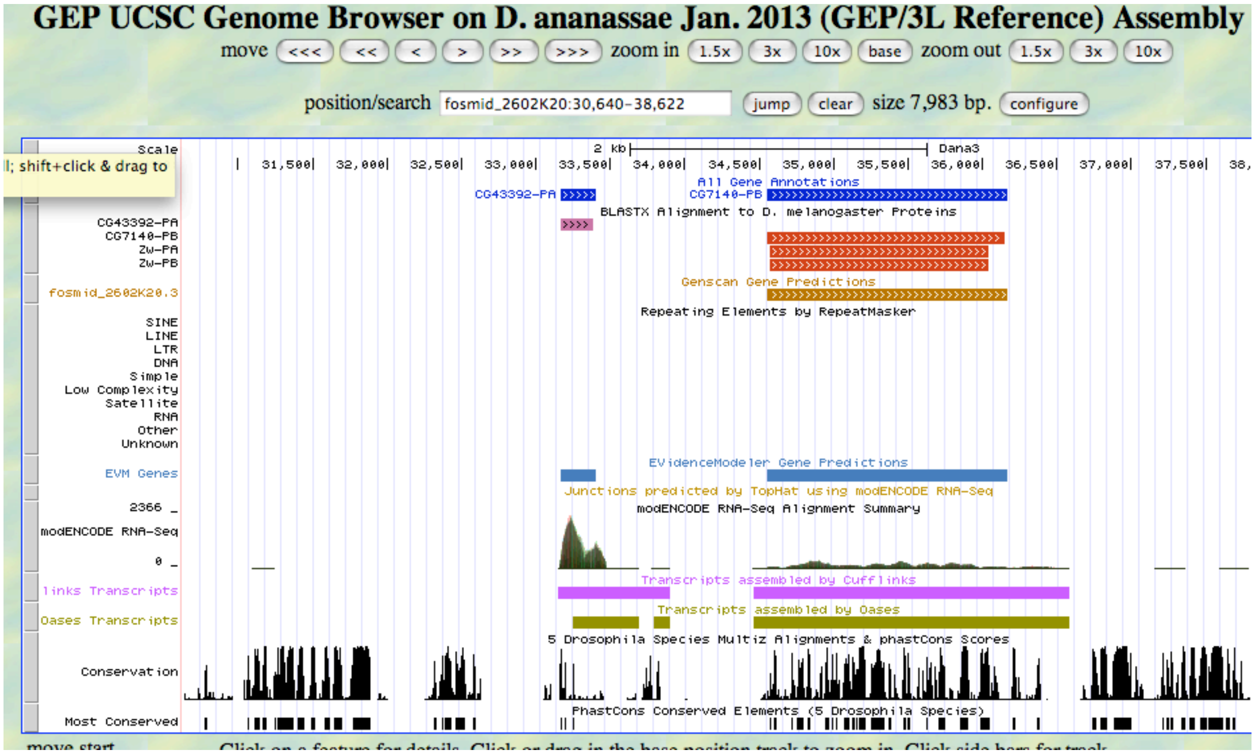
Identity: 379/536 (70.7%), **Similarity:** 457/536 (85.3%), **Gaps:** 3/536 (0.6%)

```
CG7140-PB          1 MISMDPHNEEAYSIVVFGASGGLAKKKVFPALWALFRENRLPQGTKIFTFTRSPLQTKTY 60
                     *:.:* :** *********** ***:::**.**:*:*:*******:*:** *:*****:*
Submitted_Seq      1 MLPIDLNNETAYSIVVFGASGKLAKQNIFPVLWSLYRDNRLPQGTRIYTFCRTPLQTKSY 60

CG7140-PB         61 RLQILPYMELDKHRDPKKYNLFWTTVHCVQGEYDKPENYVALTEAMVHQETKHNQVRANR 120
                     ***::***:***:*** *** ** .************ * .*.. *. :*.****. ***
Submitted_Seq     61 RLQVIPYMNLDKNRDPAKYNSFWMNVHCVQGEYDKPECYECLVSLMATHEVKHNQTVANR 120

CG7140-PB        121 IFYLALPPIVFDQVTLNVSRKCSSTTGWNRIIVEKPFARDDISYKAFQTSLCNCFRESQI 180
                     *****:*  ********.*************:*********:::::**:*****::****
Submitted_Seq    121 IFYLAVPLAVFDQVTLNTSRKCSSTTGWNRIVVEKPFARDDVTFRSFQSSLCNCYKESQI 180

CG7140-PB        181 YLMDHLLSRQVMQNFFALRYSNHLWAETLNHRHVAAVMISIKCELPVSVNRADYFNQFGI 240
                     :***:*:*:**:***:*********.****:********::*.  **...**:*******
Submitted_Seq    181 FLMDQLMSKQVLQNFLALRYSNHLWGETLNNRHVAAVMITVKSADPVPASRAEYFNQFGI 240

CG7140-PB        241 IRDLMTNHMIQMLAMLAMDQPYANTADDLRAERLKVLRQVLTPNIGDVVLAQYRNNRRET 300
                     ***:*****:**:***:*****:*..::** *******::****:***::***:**  **
Submitted_Seq    241 IRDVMTNHMVQMMAMLTMDQPYSNNPEELRIERLKVLRDILTPNMGDVIIAQYKNNGVET 300

CG7140-PB        301 DPAKCGYTEHTYIPKDSFTPTFALVVLHINNRRWTGVPFILRAGKALNDTKSEVRIQYKP 360
                     ** *******:*:**********:***:::.***:*********:****** ***:***.
Submitted_Seq    301 DPDKCGYTEHSYVPKDSFTPTFAMVVLQMKTRRWAGVPFILRAGRALNDTKVEVRVQYKA 360

CG7140-PB        361 VDCDTFHSDSTDIRNELVLRSFPTEEVFMRMRLKRQGEDICLRESEINLRVDDRGPKGLQ 420
                     .***:*****:*:*******  .* *********** ***:****:*:*********:
Submitted_Seq    361 ADCDSFHSDSSDVRNELVLRLHPFEEVFMRMRLKRPGEDLCLRETELNLRVDDRGPRVPT 420

CG7140-PB        421 GLPGYLLNVFQGDQTLFMRSDEQCEIWRIFSPVLATIDSDRPRPLHYDFGSRGPLLAYRK 480
                     .*.. **:**.*:*:****:************:*:.**.******.*:**:***: **:*
Submitted_Seq    421 SLASLLLDVFSGNQSLFMRTDEQCEIWRIFSPILSVIDNDRPRPLFYEFGTRGPIAAYHK 480

CG7140-PB        481 AERAGFVFFATDEWHQSEETLEYTVKNSKQLIGPHTALKPVRDPRSKRSN---SNP 533
                     ******:***:****** :** ***::** ******.:**:*  :*::*.   :..
Submitted_Seq    481 AERAGFLFFASDEWHQSVDTLAYTVQKSKLLIGPHTGFKPLRAIKSRKSSKRETRS 536
```

4. Dot plot between the submitted model and the *D. melanogaster ortholog*



Preparing the project for submission



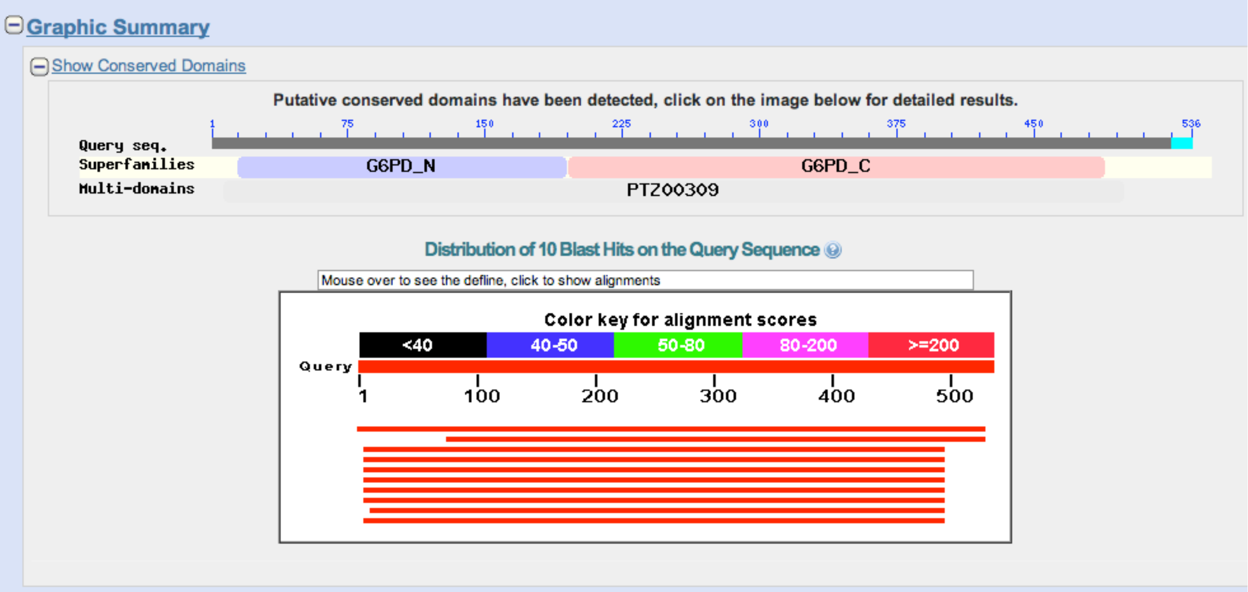
position/search fosmid_2602K20:30,640-38,622 size 7,983 bp.

ll; shift+click & drag to

move start

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track

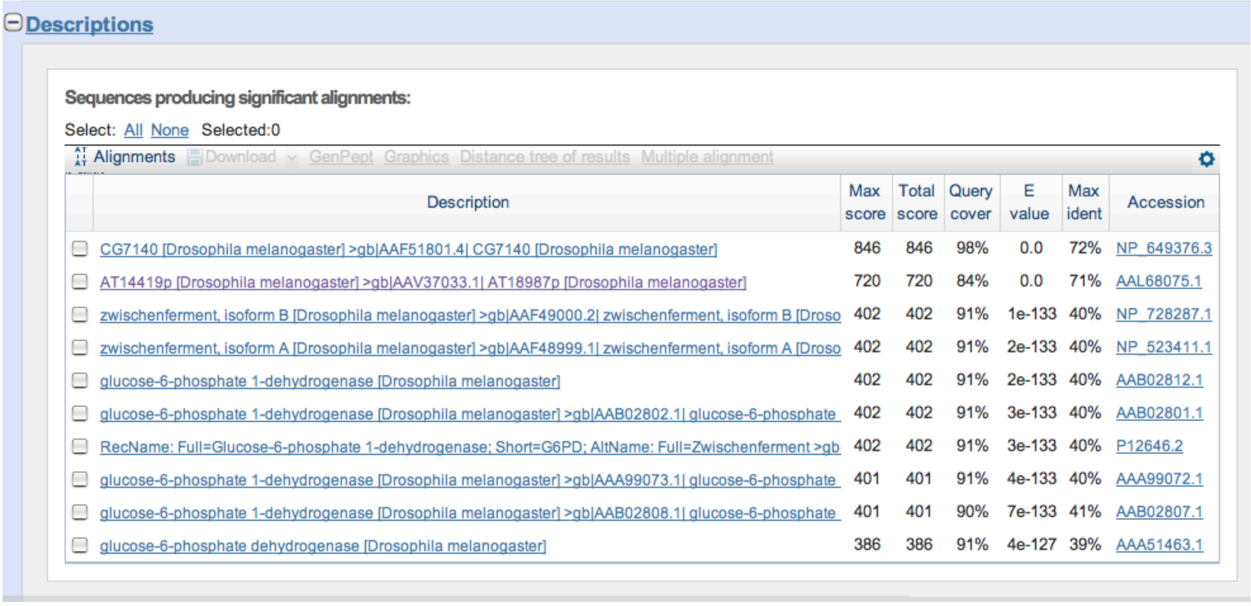
Have you annotated all the genes?



Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 10 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Sequences producing significant alignments:

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
☐	CG7140 [Drosophila melanogaster] >gb|AAF51801.4| CG7140 [Drosophila melanogaster]	846	846	98%	0.0	72%	NP_649376.3
☐	AT14419p [Drosophila melanogaster] >gb|AAV37033.1| AT18987p [Drosophila melanogaster]	720	720	84%	0.0	71%	AAL68075.1
☐	zwischenferment, isoform B [Drosophila melanogaster] >gb|AAF49000.2| zwischenferment, isoform B [Droso	402	402	91%	1e-133	40%	NP_728287.1
☐	zwischenferment, isoform A [Drosophila melanogaster] >gb|AAF48999.1| zwischenferment, isoform A [Droso	402	402	91%	2e-133	40%	NP_523411.1
☐	glucose-6-phosphate 1-dehydrogenase [Drosophila melanogaster]	402	402	91%	2e-133	40%	AAB02812.1
☐	glucose-6-phosphate 1-dehydrogenase [Drosophila melanogaster] >gb|AAB02802.1| glucose-6-phosphate	402	402	91%	3e-133	40%	AAB02801.1
☐	RecName: Full=Glucose-6-phosphate 1-dehydrogenase; Short=G6PD; AltName: Full=Zwischenferment >gb	402	402	91%	3e-133	40%	P12646.2
☐	glucose-6-phosphate 1-dehydrogenase [Drosophila melanogaster] >gb|AAA99073.1| glucose-6-phosphate	401	401	91%	4e-133	40%	AAA99072.1
☐	glucose-6-phosphate 1-dehydrogenase [Drosophila melanogaster] >gb|AAB02808.1| glucose-6-phosphate	401	401	90%	7e-133	41%	AAB02807.1
☐	glucose-6-phosphate dehydrogenase [Drosophila melanogaster]	386	386	91%	4e-127	39%	AAA51463.1

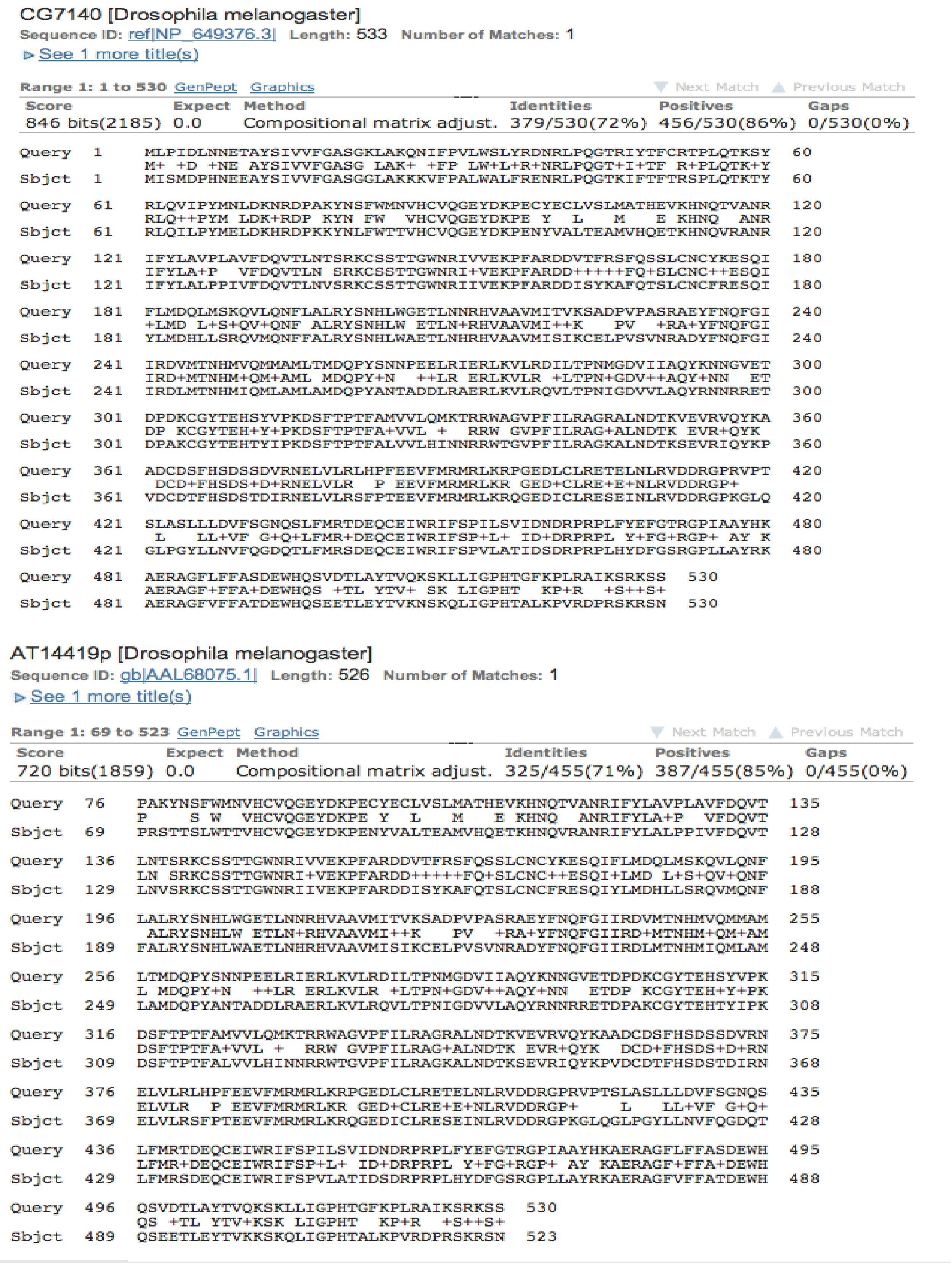
Using the E values given in the table in figure 2, I investigated the potential genes using FlyBase.

FlyBase revealed that *CG7140* is located on chromosome 3L. This is a good potential match, as we know that the fosmid came from the 3L chromosome in *D. ananassae*.

Using FlyBase to check AT14419p shows that it is an isoform of *CG43392*.

The *zwischenferment* gene (*Zw*) is located on the X chromosome and therefore it can be excluded from this analysis as a potential gene match. All of the remaining glucose-6-phosphate genes are isoforms of the *Zw* gene and can be excluded as potential matches as well.

CG3238 has a lower, yet still mildly significant alignment score. FlyBase confirmed that this gene is located on chromosome 2L. Therefore this gene can be excluded.



CG7140 [Drosophila melanogaster]

Sequence ID: ref|NP_649376.3| **Length:** 533 **Number of Matches:** 1

Range 1: 1 to 530

Score	Expect	Method	Identities	Positives	Gaps
846 bits(2185)	0.0	Compositional matrix adjust.	379/530(72%)	456/530(86%)	0/530(0%)

```
Query  1    MLPIDLNNETAYSIVVFGASGKLAKQNIFPVLWSLYRDNRLPQGTRIYTFCRTPLQTKSY  60
            M+ +D +NE AYSIVVFGASG LAK+ +FP LW+L+R+NRLPQGT+I+TF R+PLQTK+Y
Sbjct  1    MISMDPHNEEAYSIVVFGASGGLAKKKVFPALWALFRENRLPQGTKIFTFTRSPLQTKTY  60

Query  61   RLQVIPYMNLDKNRDPAKYNSFWMNVHCVQGEYDKPECYECLVSLMATHEVKHNQTVANR  120
            RLQ++PYM LDK+RDP KYN FW  VHCVQGEYDKPE Y  L   M   E KHNQ  ANR
Sbjct  61   RLQILPYMELDKHRDPKKYNLFWTTVHCVQGEYDKPENYVALTEAMVHQETKHNQVRANR  120

Query  121  IFYLAVPLAVFDQVTLNTSRKCSSTTGWNRIVVEKPFARDDVTFRSFQSSLCNCYKESQI  180
            IFYLA+P  VFDQVTLN SRKCSSTTGWNRI+VEKPFARDD+++++FQ+SLCNC++ESQI
Sbjct  121  IFYLALPPIVFDQVTLNVSRKCSSTTGWNRIIVEKPFARDDISYKAFQTSLCNCFRESQI  180

Query  181  FLMDQLMSKQVLQNFLALRYSNHLWGETLNNRHVAAVMITVKSADPVPASRAEYFNQFGI  240
            +LMD L+S+QV+QNF ALRYSNHLW ETLN+RHVAAVMI++K   PV  +RA+YFNQFGI
Sbjct  181  YLMDHLLSRQVMQNFFALRYSNHLWAETLNHRHVAAVMISIKCELPVSVNRADYFNQFGI  240

Query  241  IRDVMTNHMVQMMAMLTMDQPYSNNPEELRIERLKVLRDILTPNMGDVIIAQYKNNGVET  300
            IRD+MTNHM+QM+AML MDQPY+N  ++LR ERLKVLR +LTPN+GDV++AQY+NN  ET
Sbjct  241  IRDLMTNHMIQMLAMLAMDQPYANTADDLRAERLKVLRQVLTPNIGDVVLAQYRNNRRET  300

Query  301  DPDKCGYTEHSYVPKDSFTPTFAMVVLQMKTRRWAGVPFILRAGRALNDTKVEVRVQYKA  360
            DP KCGYTEH+Y+PKDSFTPTFA+VVL +  RRW GVPFILRAG+ALNDTK EVR+QYK
Sbjct  301  DPAKCGYTEHTYIPKDSFTPTFALVVLHINNRRWTGVPFILRAGKALNDTKSEVRIQYKP  360

Query  361  ADCDSFHSDSSDVRNELVLRLHPFEEVFMRMRLKRPGEDLCLRETELNLRVDDRGPRVPT  420
             DCD+FHSDS+D+RNELVLR  P EEVFMRMRLKR GED+CLRE+E+NLRVDDRGP+
Sbjct  361  VDCDTFHSDSTDIRNELVLRSFPTEEVFMRMRLKRQGEDICLRESEINLRVDDRGPKGLQ  420

Query  421  SLASLLLDVFSGNQSLFMRTDEQCEIWRIFSPILSVIDNDRPRPLFYEFGTRGPIAAYHK  480
             L   LL+VF G+Q+LFMR+DEQCEIWRIFSP+L+ ID+DRPRPL Y+FG+RGP+ AY K
Sbjct  421  GLPGYLLNVFQGDQTLFMRSDEQCEIWRIFSPVLATIDSDRPRPLHYDFGSRGPLLAYRK  480

Query  481  AERAGFLFFASDEWHQSVDTLAYTVQKSKLLIGPHTGFKPLRAIKSRKSS  530
            AERAGF+FFA+DEWHQS +TL YTV+ SK LIGPHT  KP+R  +S++S+
Sbjct  481  AERAGFVFFATDEWHQSEETLEYTVKNSKQLIGPHTALKPVRDPRSKRSN  530
```

AT14419p [Drosophila melanogaster]

Sequence ID: gb|AAL68075.1| **Length:** 526 **Number of Matches:** 1

Range 1: 69 to 523

Score	Expect	Method	Identities	Positives	Gaps
720 bits(1859)	0.0	Compositional matrix adjust.	325/455(71%)	387/455(85%)	0/455(0%)

```
Query  76   PAKYNSFWMNVHCVQGEYDKPECYECLVSLMATHEVKHNQTVANRIFYLAVPLAVFDQVT  135
            P    S W  VHCVQGEYDKPE Y  L   M   E KHNQ  ANRIFYLA+P  VFDQVT
Sbjct  69   PRSTTSLWTTVHCVQGEYDKPENYVALTEAMVHQETKHNQVRANRIFYLALPPIVFDQVT  128

Query  136  LNTSRKCSSTTGWNRIVVEKPFARDDVTFRSFQSSLCNCYKESQIFLMDQLMSKQVLQNF  195
            LN SRKCSSTTGWNRI+VEKPFARDD+++++FQ+SLCNC++ESQI+LMD L+S+QV+QNF
Sbjct  129  LNVSRKCSSTTGWNRIIVEKPFARDDISYKAFQTSLCNCFRESQIYLMDHLLSRQVMQNF  188

Query  196  LALRYSNHLWGETLNNRHVAAVMITVKSADPVPASRAEYFNQFGIIRDVMTNHMVQMMAM  255
             ALRYSNHLW ETLN+RHVAAVMI++K   PV  +RA+YFNQFGIIRD+MTNHM+QM+AM
Sbjct  189  FALRYSNHLWAETLNHRHVAAVMISIKCELPVSVNRADYFNQFGIIRDLMTNHMIQMLAM  248

Query  256  LTMDQPYSNNPEELRIERLKVLRDILTPNMGDVIIAQYKNNGVETDPDKCGYTEHSYVPK  315
            L MDQPY+N  ++LR ERLKVLR +LTPN+GDV++AQY+NN  ETDP KCGYTEH+Y+PK
Sbjct  249  LAMDQPYANTADDLRAERLKVLRQVLTPNIGDVVLAQYRNNRRETDPAKCGYTEHTYIPK  308

Query  316  DSFTPTFAMVVLQMKTRRWAGVPFILRAGRALNDTKVEVRVQYKAADCDSFHSDSSDVRN  375
            DSFTPTFA+VVL +  RRW GVPFILRAG+ALNDTK EVR+QYK  DCD+FHSDS+D+RN
Sbjct  309  DSFTPTFALVVLHINNRRWTGVPFILRAGKALNDTKSEVRIQYKPVDCDTFHSDSTDIRN  368

Query  376  ELVLRLHPFEEVFMRMRLKRPGEDLCLRETELNLRVDDRGPRVPTSLASLLLDVFSGNQS  435
            ELVLR  P EEVFMRMRLKR GED+CLRE+E+NLRVDDRGP+    L   LL+VF G+Q+
Sbjct  369  ELVLRSFPTEEVFMRMRLKRQGEDICLRESEINLRVDDRGPKGLQGLPGYLLNVFQGDQT  428

Query  436  LFMRTDEQCEIWRIFSPILSVIDNDRPRPLFYEFGTRGPIAAYHKAERAGFLFFASDEWH  495
            LFMR+DEQCEIWRIFSP+L+ ID+DRPRPL Y+FG+RGP+ AY KAERAGF+FFA+DEWH
Sbjct  429  LFMRSDEQCEIWRIFSPVLATIDSDRPRPLHYDFGSRGPLLAYRKAERAGFVFFATDEWH  488

Query  496  QSVDTLAYTVQKSKLLIGPHTGFKPLRAIKSRKSS  530
            QS +TL YTV+KSK LIGPHT  KP+R  +S++S+
Sbjct  489  QSEETLEYTVKKSKQLIGPHTALKPVRDPRSKRSN  523
```