Find-A-Gene Project Assignment

Name:	Steven Gan	Quarter:	Fall
PID:	A59020397	Course:	BGGN-213
E-mail:	digan@ucsd.edu	Instructor:	Dr. Barry Grant

Q1:

Name: CCCTC-binding factor (CTCF)

Accession: isoform 1 NP_006556.1 isoform 2 NP_001177951.1 isoform 3 NP_001350845.1

Accession:	isoform 1	isoform 2	isoform 3
	NP_006556.1	NP_001177951.1	NP_001350845.1

Species: *Homo Sapiens*

Function: DNA insulation; RNA binding; RNA splicing; DNA loop extrusion; transcriptional regulation; genome instability.

Q2:

Isoform 1 will be used for downstream analysis, as it is the longest isoform.

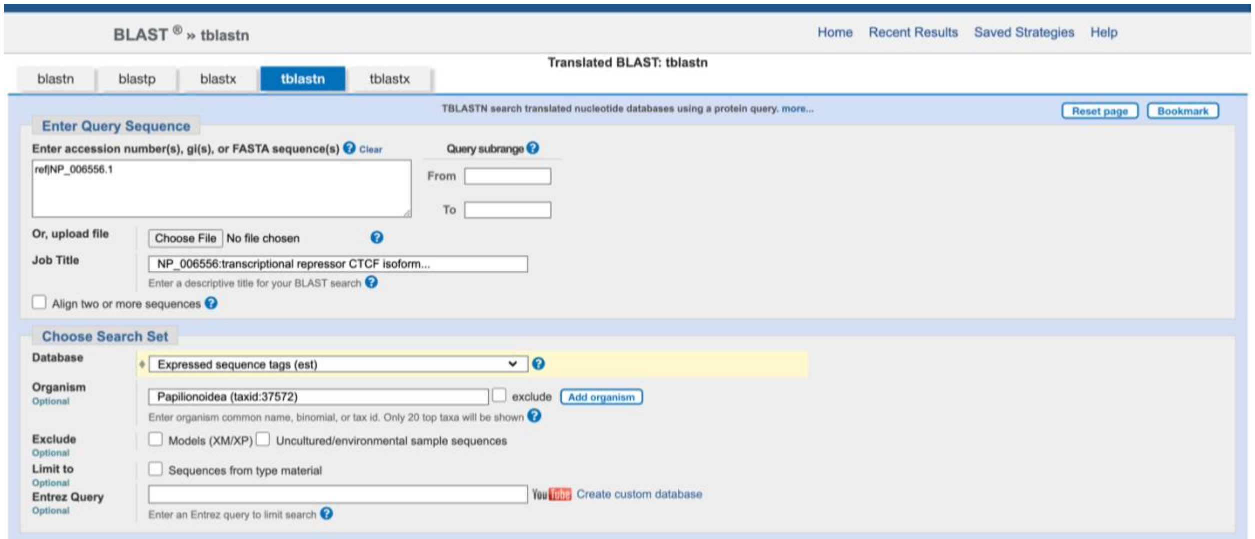
Query: NP_006556.1

Method: TBLASTN (2.13.0+) search against with default parameters

Database: Expressed Sequence Tags (est)

Organism: Papilionoidea (taxid:37572)

TBLASTN Setting:



BLAST ® » tblastn

Home | Recent Results | Saved Strategies | Help

Translated BLAST: tblastn

blastn | blastp | blastx | tblastn | tblastx

TBLASTN search translated nucleotide databases using a protein query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

ref|NP_006556.1

Query subrange
From
To

Or, upload file: Choose File No file chosen

Job Title: NP_006556:transcriptional repressor CTCF isoform...
Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Expressed sequence tags (est)

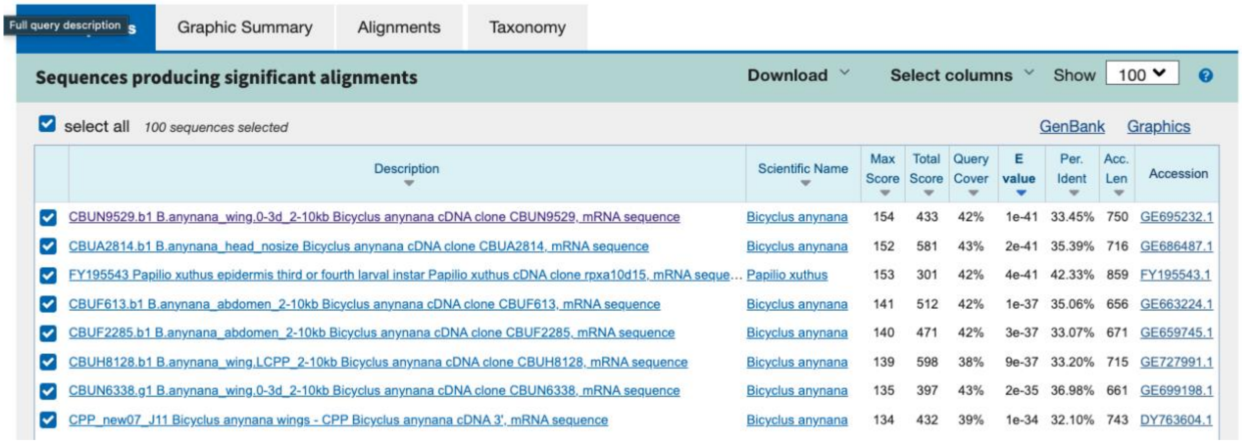
Organism (Optional): Papilionoidea (taxid:37572) □ exclude Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to (Optional): □ Sequences from type material

Entrez Query (Optional): Create custom database
Enter an Entrez query to limit search

Chosen Match: Accession GE695232.1, a 750 base pair clone from *Bicyclus anynana*. Alignment details see below.



Full query description s | Graphic Summary | Alignments | Taxonomy

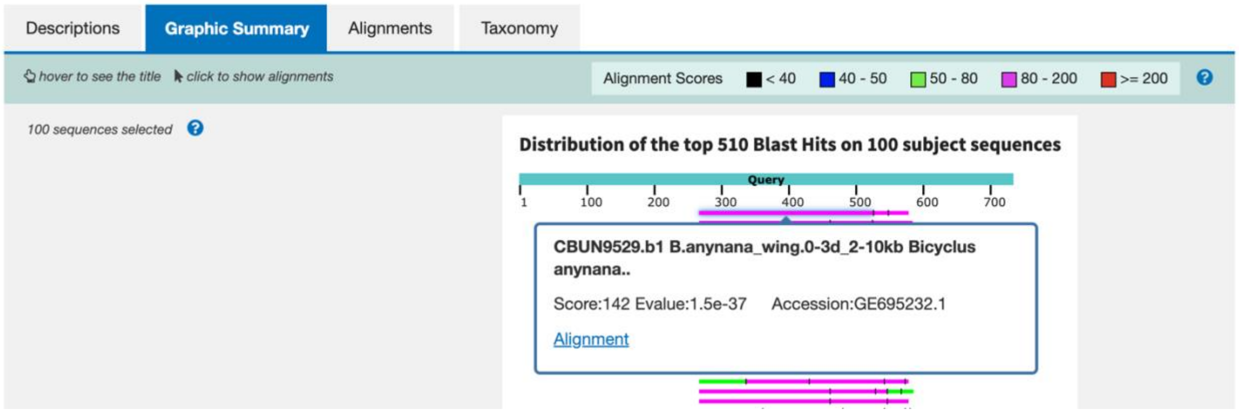
Sequences producing significant alignments

Download | Select columns | Show 100

☑ select all 100 sequences selected

GenBank | Graphics

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	CBUN9529.b1 B.anynana_wing.0-3d_2-10kb Bicyclus anynana cDNA clone CBUN9529, mRNA sequence	Bicyclus anynana	154	433	42%	1e-41	33.45%	750	GE695232.1
☑	CBUA2814.b1 B.anynana_head_nosize Bicyclus anynana cDNA clone CBUA2814, mRNA sequence	Bicyclus anynana	152	581	43%	2e-41	35.39%	716	GE686487.1
☑	FY195543 Papilio xuthus epidermis third or fourth larval instar Papilio xuthus cDNA clone rpxa10d15, mRNA seque...	Papilio xuthus	153	301	42%	4e-41	42.33%	859	FY195543.1
☑	CBUF613.b1 B.anynana_abdomen_2-10kb Bicyclus anynana cDNA clone CBUF613, mRNA sequence	Bicyclus anynana	141	512	42%	1e-37	35.06%	656	GE663224.1
☑	CBUF2285.b1 B.anynana_abdomen_2-10kb Bicyclus anynana cDNA clone CBUF2285, mRNA sequence	Bicyclus anynana	140	471	42%	3e-37	33.07%	671	GE659745.1
☑	CBUH8128.b1 B.anynana_wing.LCPP_2-10kb Bicyclus anynana cDNA clone CBUH8128, mRNA sequence	Bicyclus anynana	139	598	38%	9e-37	33.20%	715	GE727991.1
☑	CBUN6338.g1 B.anynana_wing.0-3d_2-10kb Bicyclus anynana cDNA clone CBUN6338, mRNA sequence	Bicyclus anynana	135	397	43%	2e-35	36.98%	661	GE699198.1
☑	CPP_new07_J11 Bicyclus anynana wings - CPP Bicyclus anynana cDNA 3', mRNA sequence	Bicyclus anynana	134	432	39%	1e-34	32.10%	743	DY763604.1



Descriptions | Graphic Summary | Alignments | Taxonomy

hover to see the title | click to show alignments

Alignment Scores ■ < 40 🟦 40 - 50 🟩 50 - 80 🟪 80 - 200 🟥 >= 200

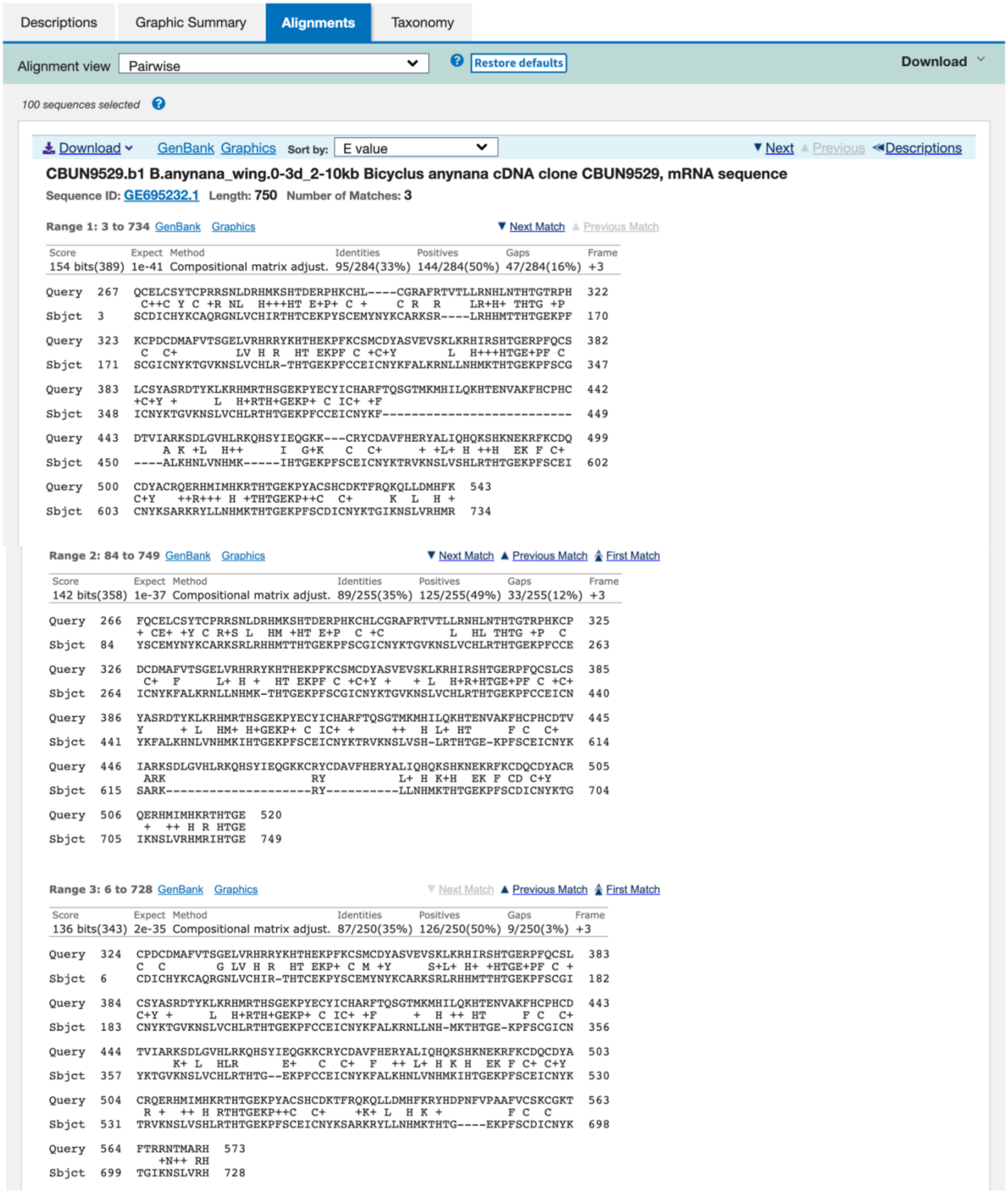
100 sequences selected

Distribution of the top 510 Blast Hits on 100 subject sequences

CBUN9529.b1 B.anynana_wing.0-3d_2-10kb Bicyclus anynana..

Score:142 Evalue:1.5e-37 Accession:GE695232.1

Alignment



CBUN9529.b1 B.anynana_wing.0-3d_2-10kb Bicyclus anynana cDNA clone CBUN9529, mRNA sequence

Sequence ID: GE695232.1 Length: 750 Number of Matches: 3

Range 1: 3 to 734 GenBank Graphics ▼ Next Match ▲ Previous Match

```
Score           Expect  Method                        Identities    Positives      Gaps          Frame
154 bits(389)   1e-41   Compositional matrix adjust.  95/284(33%)   144/284(50%)   47/284(16%)   +3

Query  267  QCELCSYTCPRRSNLDRHMKSHTDERPHKCHL----CGRAFRTVTLLRNHLNTHTGTRPH  322
             C++C Y C +R NL  H+++HT E+P+ C +    C R  R    LR+H+ THTG +P
Sbjct  3    SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSR----LRHHMTTHTGEKPF  170

Query  323  KCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCS  382
             C  C+        LV H R  HT EKPF C +C+Y       L  H+++HTGE+PF C
Sbjct  171  SCGICNYKTGVKNSLVCHLR-THTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCG  347

Query  383  LCSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHC  442
            +C+Y +     L  H+RTH+GEKP+ C IC+ +F
Sbjct  348  ICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKF--------------------------  449

Query  443  DTVIARKSDLGVHLRKQHSYIEQGKK---CRYCDAVFHERYALIQHQKSHKNEKRFKCDQ  499
                A K +L  H++     I  G+K   C  C+     + +L+ H ++H  EK F C+
Sbjct  450  ----ALKHNLVNHMK-----IHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEI  602

Query  500  CDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFK  543
            C+Y   ++R+++ H +THTGEKP++C  C+     K  L  H +
Sbjct  603  CNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLVRHMR  734
```

Range 2: 84 to 749 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

```
Score           Expect  Method                        Identities    Positives      Gaps          Frame
142 bits(358)   1e-37   Compositional matrix adjust.  89/255(35%)   125/255(49%)   33/255(12%)   +3

Query  266  FQCELCSYTCPRRSNLDRHMKSHTDERPHKCHLCGRAFRTVTLLRNHLNTHTGTRPHKCP  325
            + CE+ +Y C R+S L  HM +HT E+P  C +C         L  HL THTG +P  C
Sbjct  84   YSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCE  263

Query  326  DCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCS  385
             C+  F     L+ H +  HT EKPF C +C+Y +   + L  H+R+HTGE+PF C +C+
Sbjct  264  ICNYKFALKRNLLNHMK-THTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICN  440

Query  386  YASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTV  445
            Y     + L  HM+ H+GEKP+ C IC+ +     ++  H L+ HT     F C  C+
Sbjct  441  YKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVSH-LRTHTGE-KPFSCEICNYK  614

Query  446  IARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACR  505
             ARK                    RY          L+ H K+H  EK F CD C+Y
Sbjct  615  SARK--------------------RY----------LLNHMKTHTGEKPFSCDICNYKTG  704

Query  506  QERHMIMHKRTHTGE  520
             +  ++ H R HTGE
Sbjct  705  IKNSLVRHMRIHTGE  749
```

Range 3: 6 to 728 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

```
Score           Expect  Method                        Identities    Positives      Gaps       Frame
136 bits(343)   2e-35   Compositional matrix adjust.  87/250(35%)   126/250(50%)   9/250(3%)  +3

Query  324  CPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSL  383
            C  C       G LV H R  HT EKP+ C M +Y     S+L+ H+ +HTGE+PF C +
Sbjct  6    CDICHYKCAQRGNLVCHIR-THTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI  182

Query  384  CSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCD  443
            C+Y +     L  H+RTH+GEKP+ C IC+ +F     +  H ++ HT     F C  C+
Sbjct  183  CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNH-MKTHTGE-KPFSCGICN  356

Query  444  TVIARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYA  503
                 K+ L  HLR      E+   C  C+  F  ++ L+ H K H  EK F C+ C+Y
Sbjct  357  YKTGVKNSLVCHLRTHTG--EKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYK  530

Query  504  CRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKT  563
             R +  ++ H RTHTGEKP++C  C+    +K+ L  H K +         F C  C
Sbjct  531  TRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTG----EKPFSCDICNYK  698

Query  564  FTRRNTMARH  573
               +N++ RH
Sbjct  699  TGIKNSLVRH  728
```

Alignment Details:

CBUN9529.b1 B.anynana_wing.0-3d_2-10kb Bicyclus anynana cDNA clone CBUN9529, mRNA sequence

Sequence ID: GE695232.1   Length: 750   Number of Matches: 3

Range 1: 3 to 734

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps	Frame
154 bits (389)	1e-41	Compositional matrix adjust.	95/284(33%)	144/284(50%)	47/284(16%)	+3

```
Query  267  QCELCSYTCPRRSNLDRHMKSHTDERPHKCHL----CGRAFRTVTLLRNHLNTHTGTRPH  322
             C++C Y C +R NL  H+++HT E+P+ C +    C R  R    LR+H+ THTG +P
Sbjct  3    SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSR----LRHHMTTHTGEKPF  170

Query  323  KCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCS  382
             C  C+        LV H R  HT EKPF C +C+Y       L  H+++HTGE+PF C
Sbjct  171  SCGICNYKTGVKNSLVCHLR-THTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCG  347

Query  383  LCSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHC  442
            +C+Y +     L  H+RTH+GEKP+ C IC+ +F
Sbjct  348  ICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKF--------------------------  449

Query  443  DTVIARKSDLGVHLRKQHSYIEQGKK---CRYCDAVFHERYALIQHQKSHKNEKRFKCDQ  499
                A K +L  H++     I  G+K   C  C+     + +L+ H ++H  EK F C+
Sbjct  450  ----ALKHNLVNHMK-----IHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEI  602

Query  500  CDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFK  543
            C+Y   ++R+++ H +THTGEKP++C  C+     K  L  H +
Sbjct  603  CNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLVRHMR  734
```

Range 2: 84 to 749

Alignment statistics for match #2

Score	Expect	Method	Identities	Positives	Gaps	Frame
142 bits (358)	1e-37	Compositional matrix adjust.	89/255(35%)	125/255(49%)	33/255(12%)	+3

```
Query  266  FQCELCSYTCPRRSNLDRHMKSHTDERPHKCHLCGRAFRTVTLLRNHLNTHTGTRPHKCP  325
            + CE+ +Y C R+S L  HM +HT E+P  C +C         L  HL THTG +P  C
Sbjct  84   YSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCE  263

Query  326  DCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCS  385
             C+  F     L+ H +  HT EKPF C +C+Y +   + L  H+R+HTGE+PF C +C+
Sbjct  264  ICNYKFALKRNLLNHMK-THTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICN  440

Query  386  YASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTV  445
```

```
            Y     + L  HM+ H+GEKP+ C IC+ +     ++  H L+ HT     F C  C+
Sbjct  441  YKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVSH-LRTHTGE-KPFSCEICNYK  614

Query  446  IARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACR  505
             ARK                    RY          L+ H K+H  EK F CD C+Y
Sbjct  615  SARK--------------------RY----------LLNHMKTHTGEKPFSCDICNYKTG  704

Query  506  QERHMIMHKRTHTGE  520
             +  ++ H R HTGE
Sbjct  705  IKNSLVRHMRIHTGE  749
```

Range 3: 6 to 728

Alignment statistics for match #3

Score	Expect	Method	Identities	Positives	Gaps	Frame
136 bits (343)	2e-35	Compositional matrix adjust.	87/250(35%)	126/250(50%)	9/250(3%)	+3

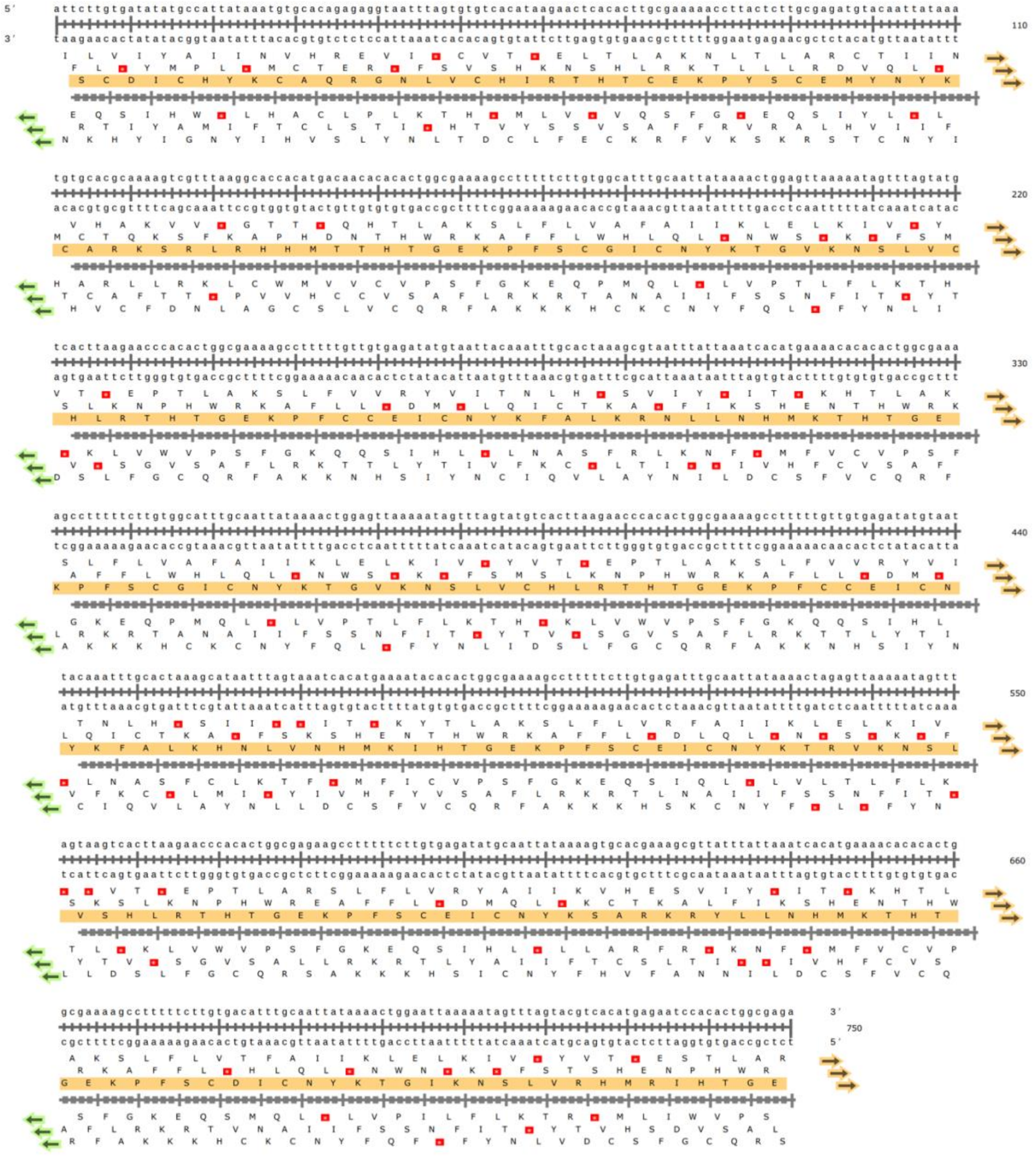
```
Query  324  CPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSL  383
            C  C       G LV H R  HT EKP+ C M +Y     S+L+ H+ +HTGE+PF C +
Sbjct  6    CDICHYKCAQRGNLVCHIR-THTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI  182

Query  384  CSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCD  443
            C+Y +     L  H+RTH+GEKP+ C IC+ +F     +  H ++ HT     F C  C+
Sbjct  183  CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNH-MKTHTGE-KPFSCGICN  356

Query  444  TVIARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYA  503
                 K+ L  HLR      E+   C  C+  F  ++ L+ H K H  EK F C+ C+Y
Sbjct  357  YKTGVKNSLVCHLRTHTG--EKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYK  530

Query  504  CRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKT  563
             R +  ++ H RTHTGEKP++C  C+    +K+ L  H K +         F C  C
Sbjct  531  TRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTG----EKPFSCDICNYK  698

Query  564  FTRRNTMARH  573
               +N++ RH
Sbjct  699  TGIKNSLVRH  728
```

Q3:

Chosen sequence:

```
>B. anynana protein (from BLAST results)
SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRT
HTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHN
LVNHMKIHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEKPFSCDICNY
KTGIKNSLVRHMRIHTG
```

All six reading frame:



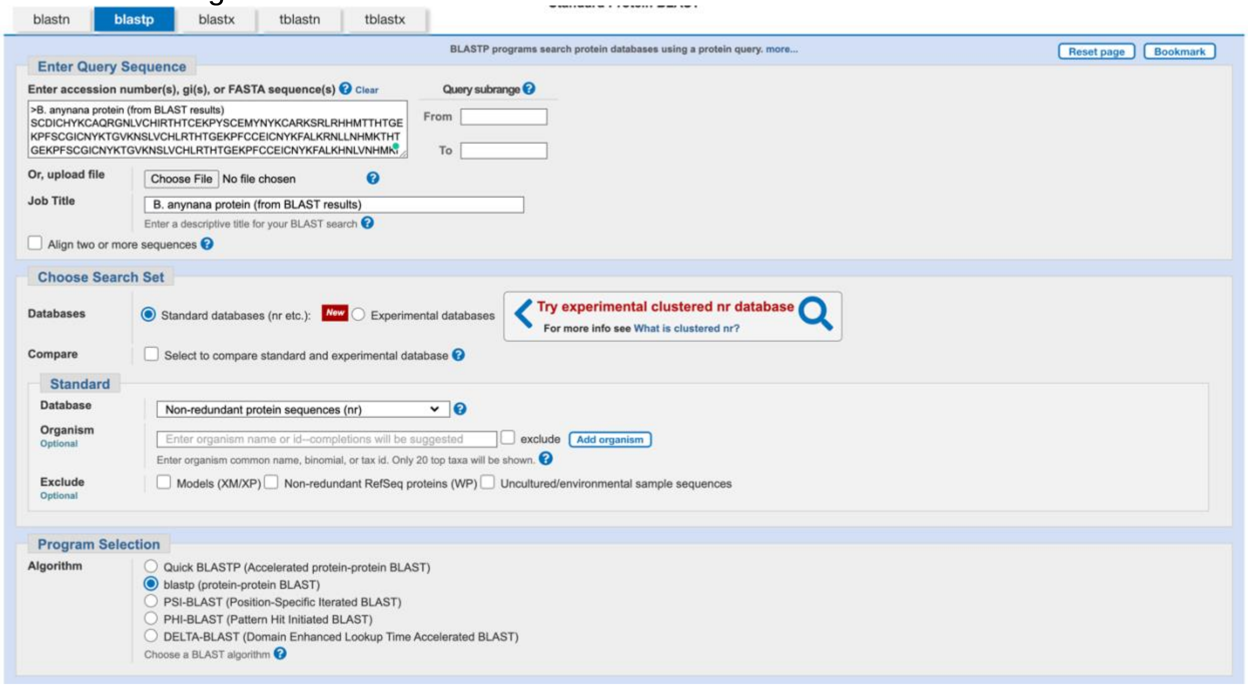
```
5' attcttgtgatatatgccattataaatgtgcacagagaggtaatttagtgtgtcacataagaactcacacttgcgaaaaaccttactcttgcgagatgtacaattataaa   110
3' taagaacactatatacggtaatatttacacgtgtctctccattaaatcacacagtgtattcttgagtgtgaacgctttttggaatgagaacgctctacatgttaatattt
   I L V I Y A I I N V H R E V I ■ C V T ■ E L T L A K N L T L A R C T I I N
   F L ■ Y M P L ■ M C T E R ■ F S V S H K N S H L R K T L L L R D V Q L ■
   S C D I C H Y K C A Q R G N L V C H I R T H T C E K P Y S C E M Y N Y K
   E Q S I H W ■ L H A C L P L K T H ■ M L V ■ V Q S F G ■ E Q S I Y L ■ L
   R T I Y A M I F T C L S T I ■ H T V Y S S V S A F F R V R A L H V I I F
   N K H Y I G N Y I H V S L Y N L T D C L F E C K R F V K S K R S T C N Y I

tgtgcacgcaaaagtcgtttaaggcaccacatgacaacacacactggcgaaaagcctttttcttgtggcatttgcaattataaaactggagttaaaaatagtttagtatg   220
acacgtgcgttttcagcaaattccgtggtgtactgttgtgtgtgaccgcttttcggaaaaagaacaccgtaaacgttaatattttgacctcaatttttatcaaatcatac
   V H A K V V ■ G T T ■ Q H T L A K S L F L V A F A I I K L E L K I V ■ Y
   M C T Q K S F K A P H D N T H W R K A F F L W H L Q L ■ N W S ■ K ■ F S M
   C A R K S R L R H H M T T H T G E K P F S C G I C N Y K T G V K N S L V C
   H A R L L R K L C W M V V C V P S F G K E Q P M Q L ■ L V P T L F L K T H
   T C A F T T ■ P V V H C C V S A F L R K R T A N A I I F S S N F I T ■ Y T
   H V C F D N L A G C S L V C Q R F A K K K H C K C N Y F Q L ■ F Y N L I

tcacttaagaacccacactggcgaaaagccttttgttgtgagatatgtaattacaaatttgcactaaagcgtaatttattaaatcacatgaaaacacacactggcgaaa   330
agtgaattcttgggtgtgaccgcttttcggaaaaacaacactctatacattaatgtttaaacgtgatttcgcattaaataatttagtgtacttttgtgtgtgaccgcttt
   V T ■ E P T L A K S L F V V R Y V I T N L H ■ S V I Y ■ I T ■ K H T L A K
   S L K N P H W R K A F L L ■ D M ■ L Q I C T K A ■ F I K S H E N T H W R K
   H L R T H T G E K P F C C E I C N Y K F A L K R N L L N H M K T H T G E
   ■ K L V W V P S F G K Q Q S I H L ■ L N A S F R L K N F ■ M F V C V P S F
   V ■ S G V S A F L R K T T L Y T I V F K C ■ L T I ■ ■ I V H F C V S A F
   D S L F G C Q R F A K K N H S I Y N C I Q V L A Y N I L D C S F V C Q R F

agcctttttcttgtggcatttgcaattataaaactggagttaaaaatagtttagtatgtcacttaagaacccacactggcgaaaagccttttgttgtgagatatgtaat   440
tcggaaaaagaacaccgtaaacgttaatattttgacctcaatttttatcaaatcatacagtgaattcttgggtgtgaccgcttttcggaaaacaacactctatacatta
   S L F L V A F A I I K L E L K I V ■ Y V T ■ E P T L A K S L F V V R Y V I
   A F F L W H L Q L ■ N W S ■ K ■ F S M S L K N P H W R K A F L L ■ D M ■
   K P F S C G I C N Y K T G V K N S L V C H L R T H T G E K P F C C E I C N
   G K E Q P M Q L ■ L V P T L F L K T H ■ K L V W V P S F G K Q Q S I H L
   L R K R T A N A I I F S S N F I T ■ Y T V ■ S G V S A F L R K T T L Y T I
   A K K K H C K C N Y F Q L ■ F Y N L I D S L F G C Q R F A K K N H S I Y N

tacaaatttgcactaaagcataatttagtaaatcacatgaaaatacacactggcgaaaagcctttttcttgtgagatttgcaattataaaactagagttaaaaatagttt   550
atgtttaaacgtgatttcgtattaaatcatttagtgtacttttatgtgtgaccgcttttcggaaaaagaacactctaaacgttaatattttgatctcaatttttatcaaa
   T N L H ■ S I I ■ ■ I T ■ K Y T L A K S L F L V R F A I I K L E L K I V
   L Q I C T K A ■ F S K S H E N T H W R K A F F L ■ D L Q L ■ N ■ S ■ K ■ F
   Y K F A L K H N L V N H M K I H T G E K P F S C E I C N Y K T R V K N S L
   ■ L N A S F C L K T F ■ M F I C V P S F G K E Q S I Q L ■ L V L T L F L K
   V F K C ■ L M I ■ Y I V H F Y V S A F L R K R T L N A I I F S S N F I T ■
   C I Q V L A Y N L L D C S F V C Q R F A K K K H S K C N Y F ■ L ■ F Y N

agtaagtcacttaagaacccacactggcgagaagcctttttcttgtgagatatgcaattataaaagtgcacgaaagcgttatttattaaatcacatgaaaacacacactg   660
tcattcagtgaattcttgggtgtgaccgctcttcggaaaaagaacactctatacgttaatattttcacgtgctttcgcaataaataatttagtgtacttttgtgtgtgac
   ■ ■ V T ■ E P T L A R S L F L V R Y A I I K V H E S V I Y ■ I T ■ K H T L
   S K S L K N P H W R E A F F L ■ D M Q L ■ K C T K A L F I K S H E N T H W
   V S H L R T H T G E K P F S C E I C N Y K S A R K R Y L L N H M K T H T
   T L ■ K L V W V P S F G K E Q S I H L ■ L L A R F R ■ K N F ■ M F V C V P
   Y T V ■ S G V S A L L R K R T L Y A I I F T C S L T I ■ ■ I V H F C V S
   L L D S L F G C Q R S A K K K H S I C N Y F H V F A N N I L D C S F V C Q

gcgaaaagcctttttcttgtgacatttgcaattataaaactggaattaaaaatagtttagtacgtcacatgagaatccacactggcgaga   3'
                                                                                          750
cgcttttcggaaaaagaacactgtaaacgttaatattttgaccttaatttttatcaaatcatgcagtgtactcttaggtgtgaccgctct   5'
   A K S L F L V T F A I I K L E L K I V ■ Y V T ■ E S T L A R
   R K A F F L ■ H L Q L ■ N W N ■ K ■ F S T S H E N P H W R
   G E K P F S C D I C N Y K T G I K N S L V R H M R I H T G E
   S F G K E Q S M Q L ■ L V P I L F L K T R ■ M L I W V P S
   A F L R K R T V N A I I F S S N F I T ■ Y T V H S D V S A L
   R F A K K K H C K C N Y F Q F ■ F Y N L V D C S F G C Q R S
```

Name:	*Bicyclus* CTCF-like protein
Species:	*Bicyclus anynana*
Taxonomy:	Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Nymphalidae; Satyrinae; Satyrini; Mycalesina; Bicyclus.

Q4:

BLASTP search on non-redundant protein sequences (nr) hits top on zinc finger protein 84-like proteins on *Bicyclus anynana*, with identity percentage of 95.16%, suggesting a possible novel protein. See details below.

BLASTP setting:



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>B. anynana protein (from BLAST results)
SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGE
KPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHT
GEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMK
```

Query subrange
From
To

Or, upload file: Choose File No file chosen

Job Title: B. anynana protein (from BLAST results)

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database
For more info see What is clustered nr?

Compare: □ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

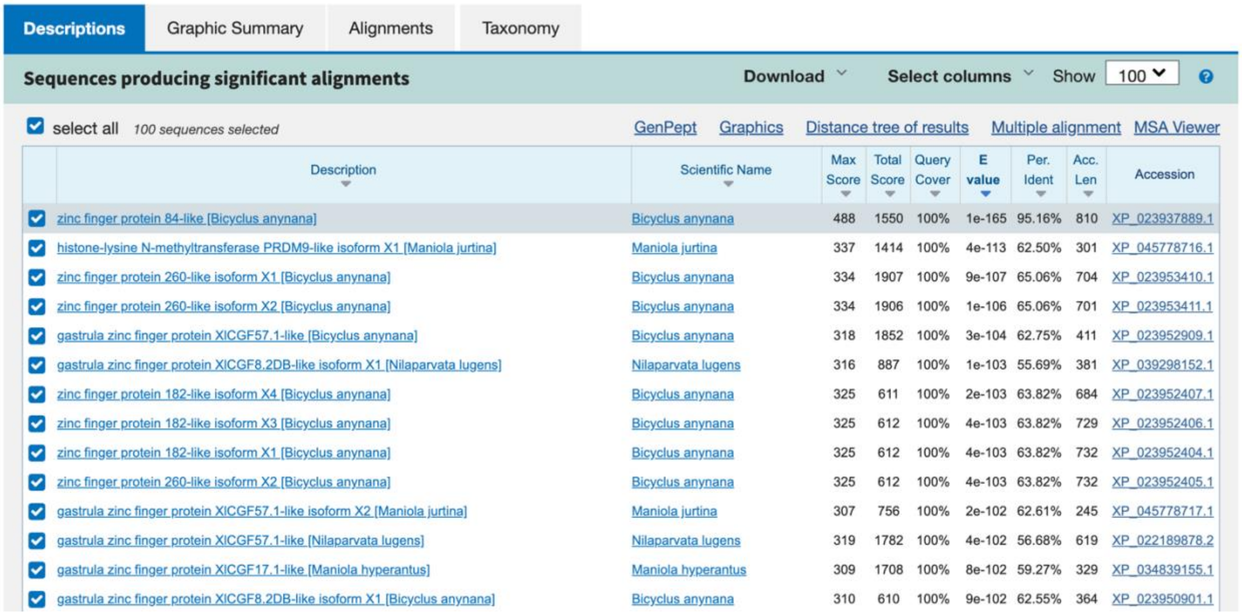
Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

Algorithm:
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLASTP results:

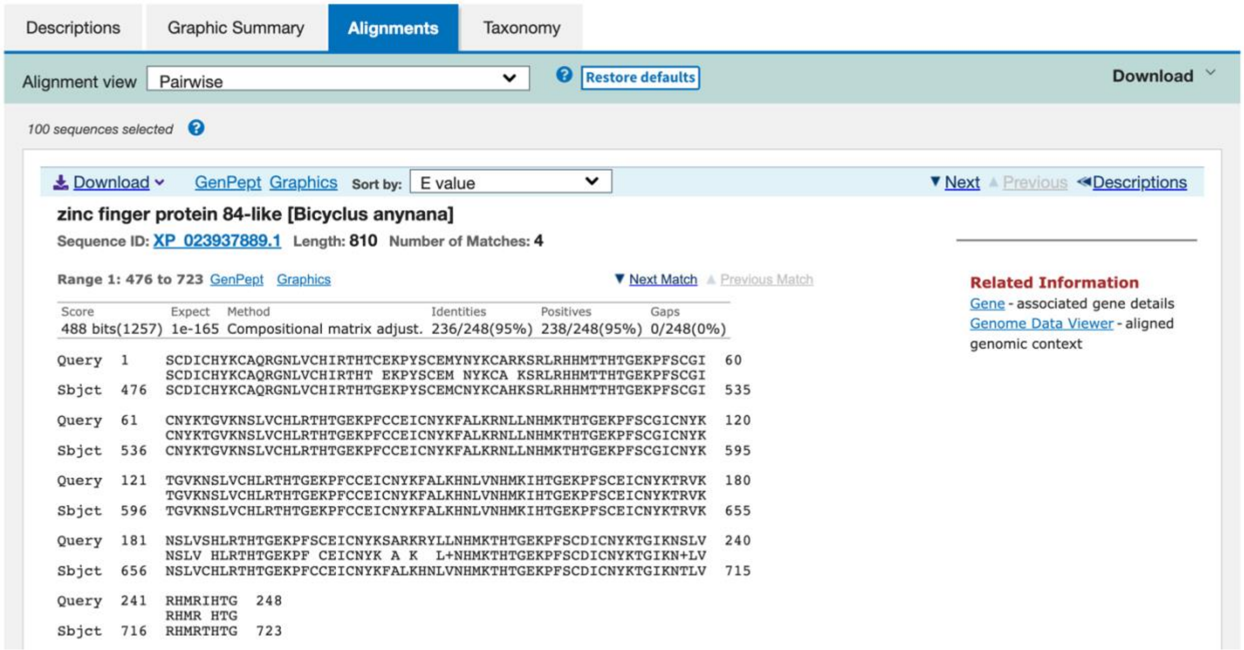


Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	zinc finger protein 84-like [Bicyclus anynana]	Bicyclus anynana	488	1550	100%	1e-165	95.16%	810	XP_023937889.1
☑	histone-lysine N-methyltransferase PRDM9-like isoform X1 [Maniola jurtina]	Maniola jurtina	337	1414	100%	4e-113	62.50%	301	XP_045778716.1
☑	zinc finger protein 260-like isoform X1 [Bicyclus anynana]	Bicyclus anynana	334	1907	100%	9e-107	65.06%	704	XP_023953410.1
☑	zinc finger protein 260-like isoform X2 [Bicyclus anynana]	Bicyclus anynana	334	1906	100%	1e-106	65.06%	701	XP_023953411.1
☑	gastrula zinc finger protein XlCGF57.1-like [Bicyclus anynana]	Bicyclus anynana	318	1852	100%	3e-104	62.75%	411	XP_023952909.1
☑	gastrula zinc finger protein XlCGF8.2DB-like isoform X1 [Nilaparvata lugens]	Nilaparvata lugens	316	887	100%	1e-103	55.69%	381	XP_039298152.1
☑	zinc finger protein 182-like isoform X4 [Bicyclus anynana]	Bicyclus anynana	325	611	100%	2e-103	63.82%	684	XP_023952407.1
☑	zinc finger protein 182-like isoform X3 [Bicyclus anynana]	Bicyclus anynana	325	612	100%	4e-103	63.82%	729	XP_023952406.1
☑	zinc finger protein 182-like isoform X1 [Bicyclus anynana]	Bicyclus anynana	325	612	100%	4e-103	63.82%	732	XP_023952404.1
☑	zinc finger protein 260-like isoform X2 [Bicyclus anynana]	Bicyclus anynana	325	612	100%	4e-103	63.82%	732	XP_023952405.1
☑	gastrula zinc finger protein XlCGF57.1-like isoform X2 [Maniola jurtina]	Maniola jurtina	307	756	100%	2e-102	62.61%	245	XP_045778717.1
☑	gastrula zinc finger protein XlCGF57.1-like [Nilaparvata lugens]	Nilaparvata lugens	319	1782	100%	4e-102	56.68%	619	XP_022189878.2
☑	gastrula zinc finger protein XlCGF17.1-like [Maniola hyperantus]	Maniola hyperantus	309	1708	100%	8e-102	59.27%	329	XP_034839155.1
☑	gastrula zinc finger protein XlCGF8.2DB-like isoform X1 [Bicyclus anynana]	Bicyclus anynana	310	610	100%	9e-102	62.55%	364	XP_023950901.1

Alignment details:



Alignment view: Pairwise

100 sequences selected

Sort by: E value

zinc finger protein 84-like [Bicyclus anynana]

Sequence ID: XP_023937889.1 Length: 810 Number of Matches: 4

Range 1: 476 to 723 GenPept Graphics

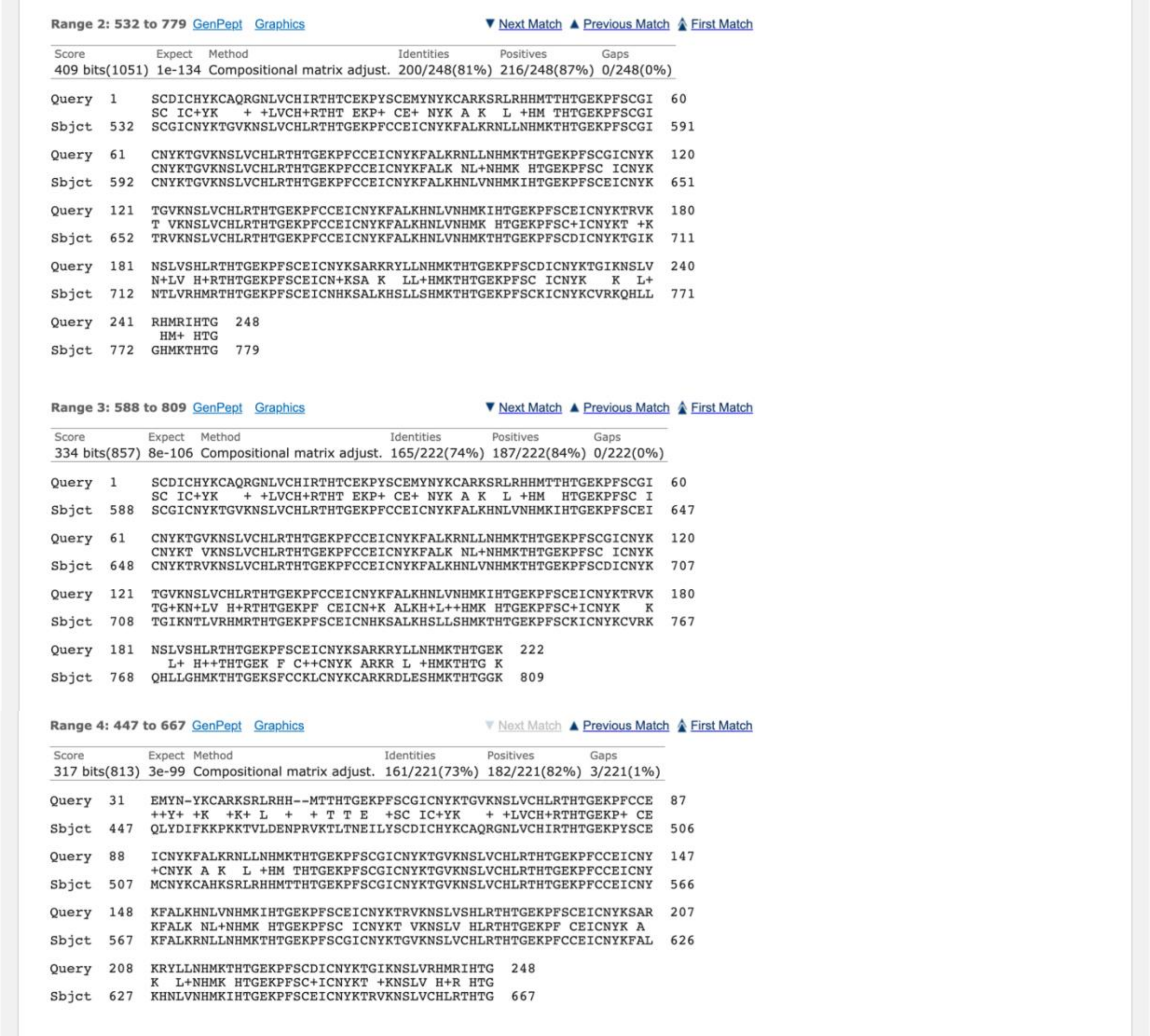
Score	Expect	Method	Identities	Positives	Gaps
488 bits(1257)	1e-165	Compositional matrix adjust.	236/248(95%)	238/248(95%)	0/248(0%)

```
Query  1    SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI  60
            SCDICHYKCAQRGNLVCHIRTHT EKPYSCEM NYKCA KSRLRHHMTTHTGEKPFSCGI
Sbjct  476  SCDICHYKCAQRGNLVCHIRTHTGEKPYSCEMCNYKCAHKSRLRHHMTTHTGEKPFSCGI  535

Query  61   CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYK  120
            CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYK
Sbjct  536  CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYK  595

Query  121  TGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVK  180
            TGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVK
Sbjct  596  TGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVK  655

Query  181  NSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLV  240
            NSLV HLRTHTGEKPF CEICNYK A K  L+NHMKTHTGEKPFSCDICNYKTGIKN+LV
Sbjct  656  NSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKTHTGEKPFSCDICNYKTGIKNTLV  715

Query  241  RHMRIHTG  248
            RHMR HTG
Sbjct  716  RHMRTHTG  723
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context



Range 2: 532 to 779 GenPept Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
409 bits(1051)	1e-134	Compositional matrix adjust.	200/248(81%)	216/248(87%)	0/248(0%)

```
Query  1     SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI  60
             SC IC+YK   + +LVCH+RTHT EKP+ CE+ NYK A K  L +HM THTGEKPFSCGI
Sbjct  532   SCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGI  591

Query  61    CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYK  120
             CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALK NL+NHMK HTGEKPFSC ICNYK
Sbjct  592   CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYK  651

Query  121   TGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVK  180
             T VKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMK HTGEKPFSC+ICNYKT +K
Sbjct  652   TRVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKTHTGEKPFSCDICNYKTGIK  711

Query  181   NSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLV  240
             N+LV H+RTHTGEKPFSCEICN+KSA K  LL+HMKTHTGEKPFSC ICNYK   K  L+
Sbjct  712   NTLVRHMRTHTGEKPFSCEICNHKSALKHSLLSHMKTHTGEKPFSCKICNYKCVRKQHLL  771

Query  241   RHMRIHTG  248
              HM+ HTG
Sbjct  772   GHMKTHTG  779
```

Range 3: 588 to 809 GenPept Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
334 bits(857)	8e-106	Compositional matrix adjust.	165/222(74%)	187/222(84%)	0/222(0%)

```
Query  1     SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI  60
             SC IC+YK   + +LVCH+RTHT EKP+ CE+ NYK A K  L +HM  HTGEKPFSC I
Sbjct  588   SCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEI  647

Query  61    CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYK  120
             CNYKT VKNSLVCHLRTHTGEKPFCCEICNYKFALK NL+NHMKTHTGEKPFSC ICNYK
Sbjct  648   CNYKTRVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKTHTGEKPFSCDICNYK  707

Query  121   TGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVK  180
             TG+KN+LV H+RTHTGEKPF CEICN+K ALKH+L++HMK HTGEKPFSC+ICNYK   K
Sbjct  708   TGIKNTLVRHMRTHTGEKPFSCEICNHKSALKHSLLSHMKTHTGEKPFSCKICNYKCVRK  767

Query  181   NSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEK  222
               L+ H++THTGEK F C++CNYK ARKR L +HMKTHTG K
Sbjct  768   QHLLGHMKTHTGEKSFCCKLCNYKCARKRDLESHMKTHTGGK  809
```

Range 4: 447 to 667 GenPept Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
317 bits(813)	3e-99	Compositional matrix adjust.	161/221(73%)	182/221(82%)	3/221(1%)

```
Query  31    EMYN-YKCARKSRLRHH--MTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCE  87
             ++Y+ +K  +K+ L  +  + T T E  +SC IC+YK   + +LVCH+RTHTGEKP+ CE
Sbjct  447   QLYDIFKKPKKTVLDENPRVKTLTNEILYSCDICHYKCAQRGNLVCHIRTHTGEKPYSCE  506

Query  88    ICNYKFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNY  147
             +CNYK A K  L +HM THTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNY
Sbjct  507   MCNYKCAHKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNY  566

Query  148   KFALKHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEICNYKSAR  207
             KFALK NL+NHMK HTGEKPFSC ICNYKT VKNSLV HLRTHTGEKPF CEICNYK A
Sbjct  567   KFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFAL  626

Query  208   KRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLVRHMRIHTG  248
             K  L+NHMK HTGEKPFSC+ICNYKT +KNSLV H+R HTG
Sbjct  627   KHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVCHLRTHTG  667
```

Q5:

Sequences for multiple alignment:

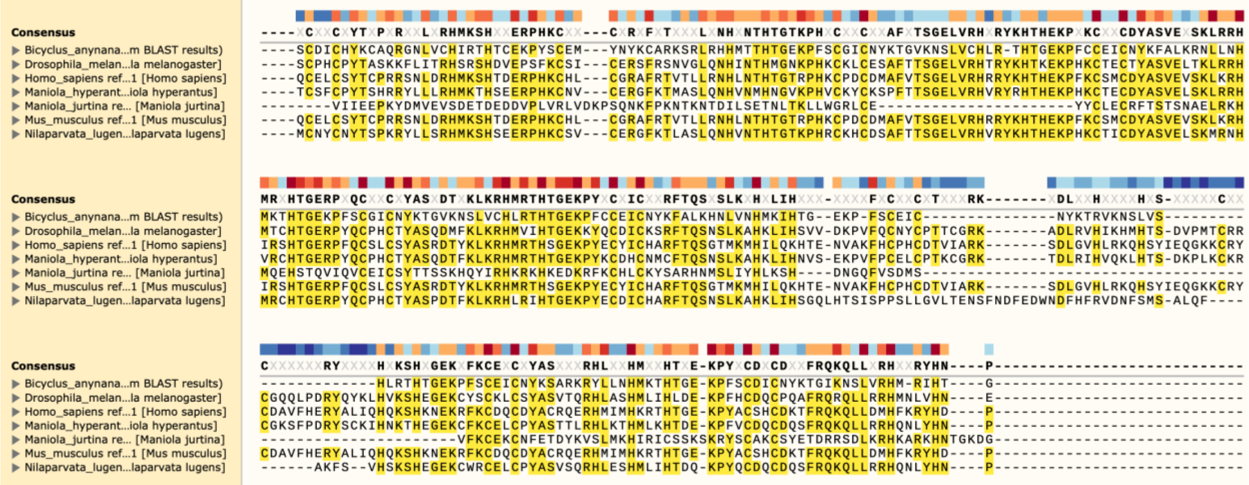
```
>Bicyclus_anynana (from BLAST results)
SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRT
HTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHN
LVNHMKIHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEKPFSCDICNY
KTGIKNSLVRHMRIHTG

>Homo_sapiens ref|NP_006556.1| transcriptional repressor CTCF isoform 1 [Homo
sapiens]
MEGDAVEAIVEESETFIKGKERKTYQRRREGGQEEDACHLPQNQTDGGEVVQDVNSSVQMVMMEQLDPTLLQMKTEV
MEGTVAPEAEAAVDDTQIITLQVVNMEEQPINIGELQLVQVPVPVTVPVATTSVEELQGAYENEVSKEGLAESEPMI
```

```
CHTLPLPEGFQVVKVGANGEVETLEQGELPPQEDPSWQKDPDYQPPAKKTKKTKKSKLRYTEEGKDVDVSVYDFEEE
QQEGLLSEVNAEKVVGNMKPPKPTKIKKKGVKKTFQCELCSYTCPRRSNLDRHMKSHTDERPHKCHLCGRAFRTVTL
LRNHLNTHTGTRPHKCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCS
YASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTVIARKSDLGVHLRKQHSY
IEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLD
MHFKRYHDPNFVPAAFVCSKCGKTFTRRNTMARHADNCAGPDGVEGENGGETKKSKRGRKRKMRSKKEDSSDSENAE
PDLDDNEDEEEPAVEIEPEPEPQPVTPAPPPAKKRRGRPPGRTNQPKQNQPTAIIQVEDQNTGAIENIIVEVKKEPD
AEPAEGEEEEAQPAATDAPNGDLTPEMILSMMDR

>Maniola_jurtina ref| XP_045782119.1 | transcriptional repressor CTCF-like
isoform X2 [Maniola jurtina]
MAGICCVDGCDPTAEDVTYFKFPNSRTLRRKWLDAINNSVKVTLDTAVCSRHFLPNQYEVIRGKKRLKAKVVPSVFD
NITKPTSPQKEKTDSSDGEDSVPLQKVKSVATDNTDTGQSKQSPDHRLEDRQDSEASVRNGVKEDDIVKRKQPDIVS
ITDSESNKDIEDIITHYQIKQIRPLHKPTDRTTPDMSIEIEVPLAMDGEMEIGRNREMEMGRNGEIDDVIDVDEEAE
PVFIEVAVGKGGGVEETTNEDCMMLLESVQCEVDPSCLMFPPEDPGNDPGDDAGNDSDVIDLGEKKEDPVSLLTSSD
EDEVIIEEPKYDMVEVSDETDEDDVPLVRLVDKPSQNKFPKNTKNTDILSETNLTKLLWGRLCEYYCLECRFTSTSN
AELRKHMQEHSTQVIQVCEICSYTTSSKHQYIRHKRKHKEDKRFKCHLCKYSARHNMSLIYHLKSHDNGQFVSDMSV
FKCEKCNFETDYKVSLMKHIRICSSKSKRYSCAKCSYETDRRSDLKRHKARKHNTGKDGDYEPPAWVTRAKKPKCDK

>Nilaparvata_lugens ref|XP_039287301.1| transcriptional repressor CTCF
[Nilaparvata lugens]
MSPPDKVQVQTEIKLEDGVTIVPENVTDIQNYLDTFNKEIQGGEQVVQQVGVVAADEGGSEEGTYYVDQAGQYYYQS
ASCDGQQVMTVVSGLPGASGESGESFVALPASAASSQRDNVGGSAPLLIQAATGGGASASGGGTTVGGATVGGASAE
GGGGATYQTVTIVPSETNPGELSYLLIVQQPGDEDEGEDGQKDKDEDDDHDLTVYDFDDAEDVGTVSGMESGDEDDK
SKIVKFMPKKSQTVTQAHMCNYCNYTSPKRYLLSRHMKSHSEERPHKCSVCERGFKTLASLQNHVNTHTGTKPHRCK
HCDSAFTTSGELVRHVRYKHTHEKPHKCTICDYASVELSKMRNHMRCHTGERPYQCPHCTYASPDTFKLKRHLRIHT
GEKPYECDICHARFTQSNSLKAHKLIHSGQLHTSISPPSLLGVLTENSFNDFEDWNDFHFRVDNFSMSALQFAKFSV
HSKSHEGEKCWRCELCPYASVSQRHLESHMLIHTDQKPYQCDQCDQSFRQKQLLRRHQNLYHNPNYVPPPPREKTHE
CPECQRAFRHKGNLIRHLSVHDPESLAQERQMLLKQGRQRKLQNINGQRVEVIPGDEDDEDEDDELNGQVMAVEGSD
GQQYVVLEVIQLQDDNGQEQAVAVMAADGGLEQAVAALHGAGAEDDEEELDEEDEEEDDVHITPDMVEDHEMMSSLR
HQTRQKTQHDMANCFGFDDDEEEEEEDDIGISLKQSNTKSIHLLRSGLQ

>Mus_musculus ref|NP_001390652.1| transcriptional repressor CTCF isoform 1
[Mus musculus]
MEGEAVEAIVEESETFIKGKERKTYQRRREGGQEEDACHLPQNQTDGGEVVQDVNSSVQMVMMEQLDPTLLQMKTEV
MEGTVAPEAEAAVDDTQIITLQVVNMEEQPINIGELQLVQVPVPVTVPVATTSVEELQGAYENEVSKEGLAESEPMI
CHTLPLPEGFQVVKVGANGEVETLEQGELPPQEDSSWQKDPDYQPPAKKTKKTKKSKLRYTEEGKDVDVSVYDFEEE
QQEGLLSEVNAEKVVGNMKPPKPTKIKKKGVKKTFQCELCSYTCPRRSNLDRHMKSHTDERPHKCHLCGRAFRTVTL
LRNHLNTHTGTRPHKCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCS
YASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTVIARKSDLGVHLRKQHSY
IEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLD
MHFKRYHDPNFVPAAFVCSKCGKTFTRRNTMARHADNCAGPDGVEGENGGETKKSKRGRKRKMRSKKEDSSDSEENA
EPDLDDNEEEEEPAVEIEPEPEPQPQPPPPPQPVAPAPPPAKKRRGRPPGRTNQPKQNQPTAIIQVEDQNTGAIENI
IVEVKKEPDAEPAEGEEEEAQAATTDAPNGDLTPEMILSMMDR

>Drosophila_melanogaster ref|NP_648109.1| CTCF [Drosophila melanogaster]
MPRRTKKDEDPEDLQTFLNNFHKEIEGNSDEKVVNTILEAISAEAIDLDENGAEAGGSKPMEEAEADLDHAEEAEEE
EEDDEDKYFIDDEGNCYIKTTPKKQKELQKKLKQAAAKPGKATRSVVSTATNKSINLRPAKSTPKATTSKPPPEPKA
ISVRPARAAAAKAKQSAMPPPPALVVKVPAPRGRPRKNPVIPKPEPMDLERELEELVDEPDISSMVTELSDYTVDEA
AVEAATATLTPNEAEVYEFEDNATTEDENADKKDVDFVLSNKEVKLKTASSTSQNSNASGHKYSCPHCPYTASKKFL
ITRHSRSHDVEPSFKCSICERSFRSNVGLQNHINTHMGNKPHKCKLCESAFTTSGELVRHTRYKHTKEKPHKCTECT
YASVELTKLRRHMTCHTGERPYQCPHCTYASQDMFKLKRHMVIHTGEKKYQCDICKSRFTQSNSLKAHKLIHSVVDK
PVFQCNYCPTTCGRKADLRVHIKHMHTSDVPMTCRRCGQQLPDRYQYKLHVKSHEGEKCYSCKLCSYASVTQRHLAS
HMLIHLDEKPFHCDQCPQAFRQRQLLRRHMNLVHNEEYQPPEPREKLHKCPSCPREFTHKGNLMRHMETHDDSANAR
EKRRRLKLGRNVRLQKDGTVITLIKDQYVDMDRDQEENEEDDNPESYDLAEIEPENSEAEDADDDVETIVSDPIRQR
IKPAPIIINKQARLAASEKQPMIINQRLRSQRGTKTFHIKEEPDNSDFTVEWQGDDGEVMVVELVNGDEEVLVKHEP
```

```
SANSKISAKNCFGFEDDDDYEEYGDGENEVDGASQEFLQLMDMIEQDS

>Maniola_hyperantus ref|XP_034839420.1| transcriptional repressor CTCF-like
[Maniola hyperantus]
MPPPDKKSANKCKTILQTYLNSFDQDNEPTTVIVNGDGDEADAGVTYFVDEEGRYYYQPAGDSQNLVSLPIEAEAED
GTEIPQEAQMLVDGDGYQTVTLMPSEEGGELSYVLVMQEETKPVMNIDIKVDQDEEKSSDVYKFEEEEEEDPPIEVS
DEVEESIKPKLTFAMKRSKHLRPSFTCSFCPYTSHRRYLLLRHMKTHSEERPHKCNVCERGFKTMASLQNHVNMHNG
VKPHVCKYCKSPFTTSGELVRHVRYRHTHEKPHKCTECDYASVELSKLRRHVRCHTGERPYQCPHCTYASQDTFKLK
RHMRTHTGEKPYKCDHCNMCFTQSNSLKAHKLIHNVSEKPVFPCELCPTKCGRKTDLRIHVQKLHTSDKPLKCKRCG
KSFPDRYSCKIHNKTHEGEKCFKCELCPYASTTLRHLKTHMLKHTDEKPFVCDQCDQSFRQKQLLRRHQNLYHNPNY
EPKPPKEKTHTCHECKRTFAHKGNLIRHLAIHDPDSGHQERALALRLGRQKKIKFVDGNVKTDDSDNEPEEIMKLDL
GGNQLERGELLTVADNDGQQYVVLEVIQAEDGETQIVSAADYEEEEEEEEEEEDEDDEELDKKEIIYEQIKPKGMME
RTIKLESDVDTCFGFDEDEEEPDEDEEGIAYNDKIVLRIV
```

Alignment:

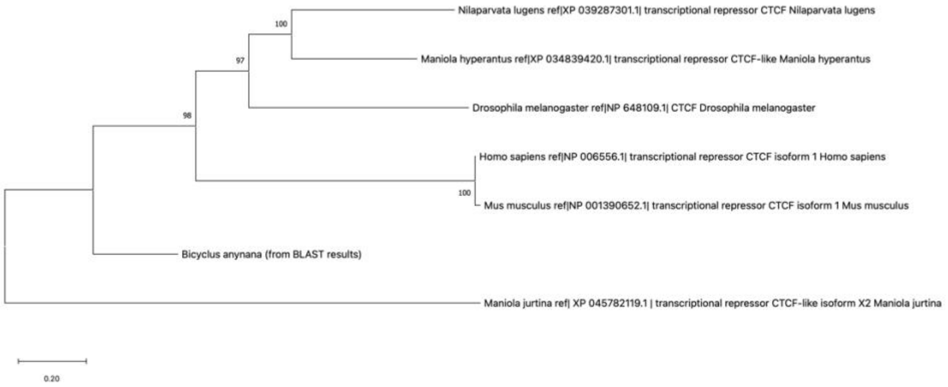
Obtained using MUSCLE (version 3.8) in SnapGene (version 6.0.2)



```
Consensus                               ----XCXXCXYTXPXRXXLXRHMKSHXXERPHKCXX---CXRXFXTXXXLXNHXNTHTGTKPHXCXXCXXAFXTSGELVRHXRYKHTHEKPXKCXXCDYASVEXSKLRRH
▶ Bicyclus_anynana...m BLAST results)   ----SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEM---YNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLR-THTGEKPFCCEICNYKFALKRNLLNH
▶ Drosophila_melan...la melanogaster]   ----SCPHCPYTASKKFLITRHSRSHDVEPSFKCSI---CERSFRSNVGLQNHINTHMGNKPHKCKLCESAFTTSGELVRHTRYKHTKEKPHKCTECTYASVELTKLRRH
▶ Homo_sapiens ref...1 [Homo sapiens]   ----QCELCSYTCPRRSNLDRHMKSHTDERPHKCHL---CGRAFRTVTLLRNHLNTHTGTRPHKCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRH
▶ Maniola_hyperant...iola hyperantus]   ----TCSFCPYTSHRRYLLLRHMKTHSEERPHKCNV---CERGFKTMASLQNHVNMHNGVKPHVCKYCKSPFTTSGELVRHVRYRHTHEKPHKCTECDYASVELSKLRRH
▶ Maniola_jurtina re... [Maniola jurtina] --------VIIEEPKYDMVEVSDETDEDDVPLVRLVDKPSQNKFPKNTKNTDILSETNLTKLLWGRLCE----------------------YYCLECRFTSTSNAELRKH
▶ Mus_musculus ref...1 [Mus musculus]   ----QCELCSYTCPRRSNLDRHMKSHTDERPHKCHL---CGRAFRTVTLLRNHLNTHTGTRPHKCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRH
▶ Nilaparvata_lugen...laparvata lugens] ----MCNYCNYTSPKRYLLSRHMKSHSEERPHKCSV---CERGFKTLASLQNHVNTHTGTKPHRCKHCDSAFTTSGELVRHVRYKHTHEKPHKCTICDYASVELSKMRNH

Consensus                               MRXHTGERPXQCXXCXYASXDTXKLKRHMRTHTGEKPYXCXICXXRFTQSXSLKXHXLIHXXX-XXXXFXCXXXCXTXXXRK-------XDLXXHXXXXHXS-XXXXXCXX
▶ Bicyclus_anynana...m BLAST results)   MKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTG--EKP-FSCEIC---------------NYKTRVKNSLVS---------
▶ Drosophila_melan...la melanogaster]   MTCHTGERPYQCPHCTYASQDMFKLKRHMVIHTGEKKYQCDICKSRFTQSNSLKAHKLIHSVV-DKPVFQCNYCPTTCGRK-------ADLRVHIKHMHTS-DVPMTCRR
▶ Homo_sapiens ref...1 [Homo sapiens]   IRSHTGERPFQCSLCSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTE-NVAKFHCPHCDTVIARK-------SDLGVHLRKQHSYIEQGKKCRY
▶ Maniola_hyperant...iola hyperantus]   VRCHTGERPYQCPHCTYASQDTFKLKRHMRTHTGEKPYKCDHCNMCFTQSNSLKAHKLIHNVS-EKPVFPCELCPTKCGRK-------TDLRIHVQKLHTS-DKPLKCKR
▶ Maniola_jurtina re... [Maniola jurtina] MQEHSTQVIQVCEICSYTTSSKHQYIRHKRKHKEDKRFKCHLCKYSARHNMSLIYHLKSH----DNGQFVSDMS------------------------------------
▶ Mus_musculus ref...1 [Mus musculus]   IRSHTGERPFQCSLCSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTE-NVAKFHCPHCDTVIARK-------SDLGVHLRKQHSYIEQGKKCRY
▶ Nilaparvata_lugen...laparvata lugens] MRCHTGERPYQCPHCTYASPDTFKLKRHLRIHTGEKPYECDICHARFTQSNSLKAHKLIHSGQLHTSISPPSLLGVLTENSFNDFEDWNDFHFRVDNFSMS-ALQF----

Consensus                               CXXXXXXRYXXXXHXKSHXGEKXFKCEXCXYASXXXRHLXXHMXXHTXE-KPYXCDXCDXXFRQKQLLXRHXXRYHN----P----------------------------
▶ Bicyclus_anynana...m BLAST results)   -------------HLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGE-KPFSCDICNYKTGIKNSLVRHM-RIHT----G----------------------------
▶ Drosophila_melan...la melanogaster]   CGQQLPDRYQYKLHVKSHEGEKCYSCKLCSYASVTQRHLASHMLIHLDE-KPFHCDQCPQAFRQRQLLRRHMNLVHN----E----------------------------
▶ Homo_sapiens ref...1 [Homo sapiens]   CDAVFHERYALIQHQKSHKNEKRFKCDQCDYACRQERHMIMHKRTHTGE-KPYACSHCDKTFRQKQLLDMHFKRYHD----P----------------------------
▶ Maniola_hyperant...iola hyperantus]   CGKSFPDRYSCKIHNKTHEGEKCFKCELCPYASTTLRHLKTHMLKHTDE-KPFVCDQCDQSFRQKQLLRRHQNLYHN----P----------------------------
▶ Maniola_jurtina re... [Maniola jurtina] ----------------------VFKCEKCNFETDYKVSLMKHIRICSSKSKRYSCAKCSYETDRRSDLKRHKARKHNTGKDG----------------------------
▶ Mus_musculus ref...1 [Mus musculus]   CDAVFHERYALIQHQKSHKNEKRFKCDQCDYACRQERHMIMHKRTHTGE-KPYACSHCDKTFRQKQLLDMHFKRYHD----P----------------------------
▶ Nilaparvata_lugen...laparvata lugens] ------AKFS--VHSKSHEGEKCWRCELCPYASVSQRHLESHMLIHTDQ-KPYQCDQCDQSFRQKQLLRRHQNLYHN----P----------------------------
```

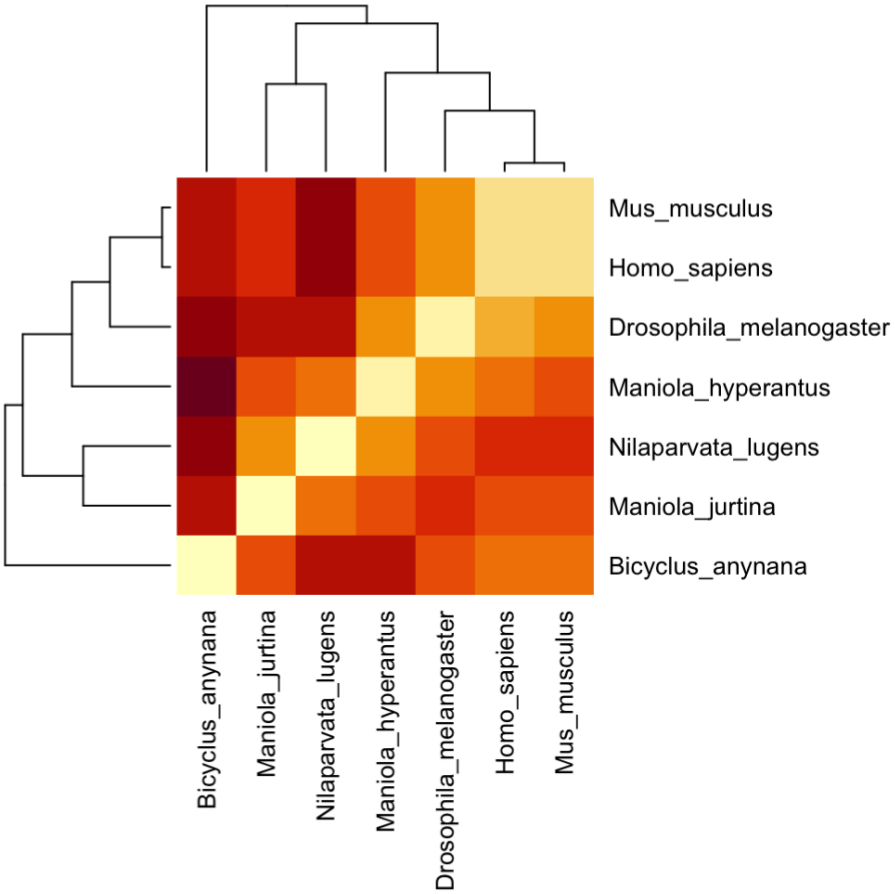
Q6:

Phylogenetic tree: Generated using MEGA (11.0.13), aligned with MUSCLE (3.8)


Nilaparvata lugens ref|XP 039287301.1| transcriptional repressor CTCF Nilaparvata lugens
100
Maniola hyperantus ref|XP 034839420.1| transcriptional repressor CTCF-like Maniola hyperantus
97
Drosophila melanogaster ref|NP 648109.1| CTCF Drosophila melanogaster
98
Homo sapiens ref|NP 006556.1| transcriptional repressor CTCF isoform 1 Homo sapiens
100
Mus musculus ref|NP 001390652.1| transcriptional repressor CTCF isoform 1 Mus musculus
Bicyclus anynana (from BLAST results)
Maniola jurtina ref| XP 045782119.1 | transcriptional repressor CTCF-like isoform X2 Maniola jurtina
0.20

Q7:

Heatmap:



Q8:

Top three hits of blast search based on the consensus sequence:

ID	Technique	Resolution	Source	E-value	Identity (%)
6QNX_C	X-ray Diffraction	2.700	*Homo sapiens*	1.21e-95	56.584
5YEF_A	X-ray Diffraction	2.807	*Homo sapiens*	5.49e-66	57.071
5T0U_A	X-ray Diffraction	3.199	*Homo sapiens*	1.32e-62	61.677

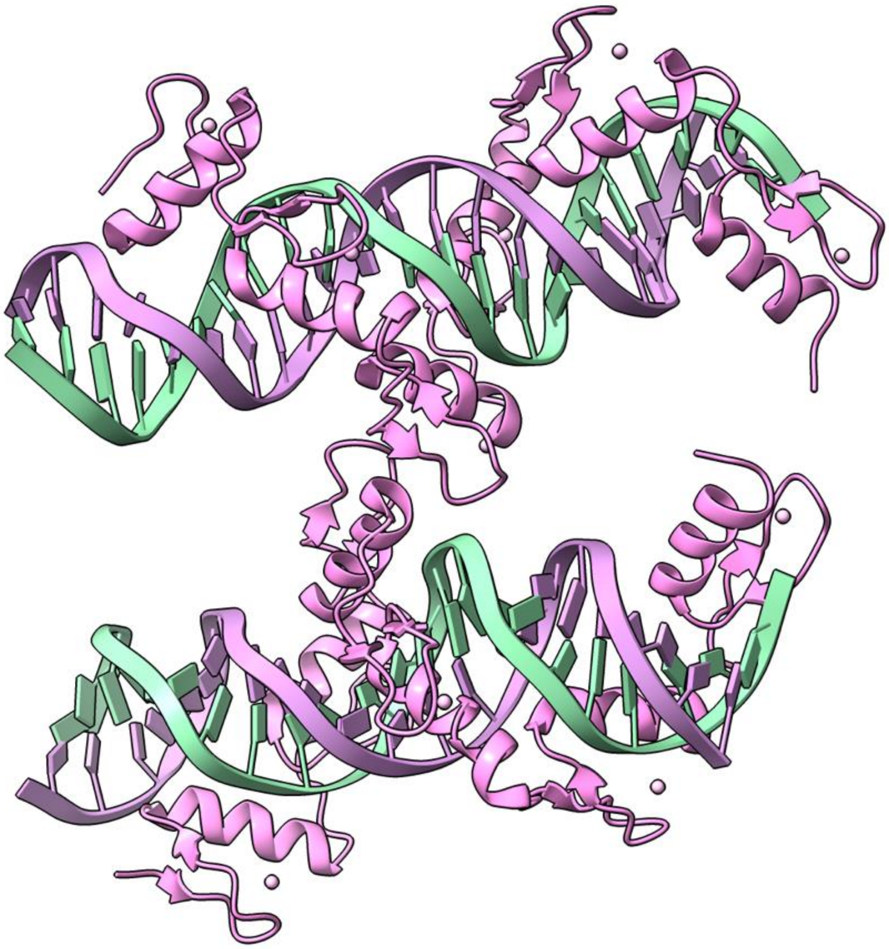
Consensus sequence:

```
>Aligned Consensus (threshold of >50%)
XCXXCXYTXPXRXXLXRHMKSHXXERPHKCXXCXRXFXTXXXLXNHXNTHTGTKPHXCXXCXXAFXTSGELVRHXRY
KHTHEKPXKCXXCDYASVEXSKLRRHMRXHTGERPXQCXXCXYASXDTXKLKRHMRTHTGEKPYXCXICXXRFTQSX
SLKXHXLIHXXXXXXXFXCXXCXTXXXRKXDLXXHXXXXHXSXXXXXCXXCXXXXXXRYXXXXHXKSHXGEKXFKCE
XCXYASXXXRHLXXHMXXHTXEKPYXCDXCDXXFRQKQLLXRHXXRYHNP
```

Q9:

5T0U might not bear too much similarity to *Bicyclus anynana* CTCF like proteins since the identity only go as high as 62%. Figure below is the structure of homodimer CTCF proteins in human binding to two DNA strands.

Structure of 5T0U:



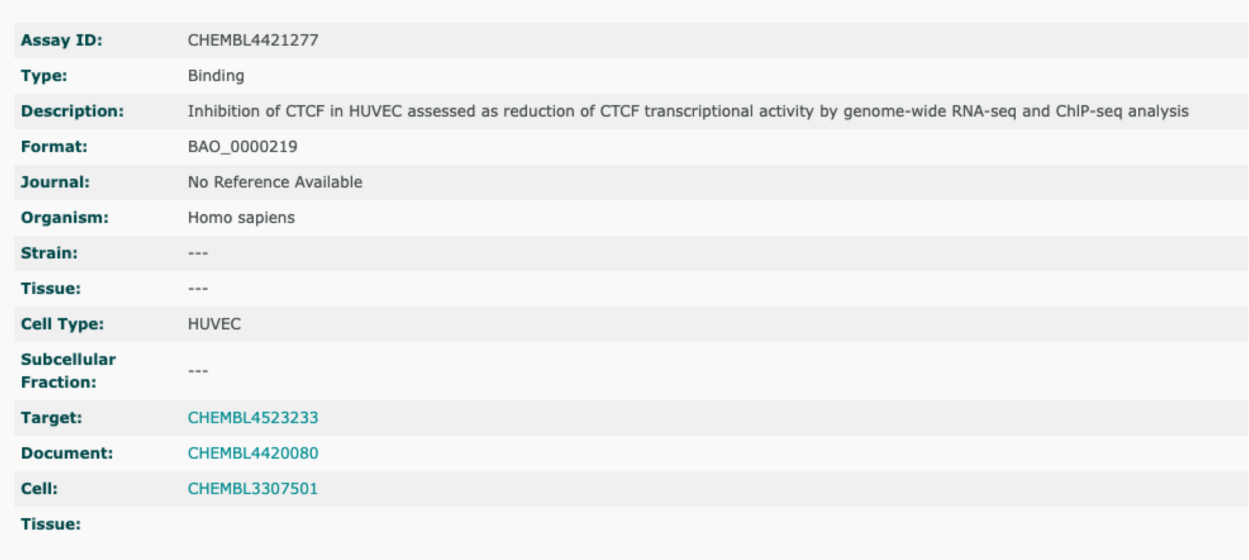
Q10:

https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL4523233/

Target report of CTCF protein in ChEMBL:

Only one chemical, CHEMBL2334661 ($C_{19}H_{16}O_3$; name undefined), is reported to inhibit CTCF in HUVEC (*Homo Sapiens* cell line) by reducing *CTCF* gene transcription.

Assay ID: CHEMBL4421277



Assay ID:	CHEMBL4421277
Type:	Binding
Description:	Inhibition of CTCF in HUVEC assessed as reduction of CTCF transcriptional activity by genome-wide RNA-seq and ChIP-seq analysis
Format:	BAO_0000219
Journal:	No Reference Available
Organism:	Homo sapiens
Strain:	---
Tissue:	---
Cell Type:	HUVEC
Subcellular Fraction:	---
Target:	CHEMBL4523233
Document:	CHEMBL4420080
Cell:	CHEMBL3307501
Tissue:	