Sequence Updater User Guide

Wilson Leung

Table of Contents

Introduction

The genome assemblies used by the GEP scientific projects have not undergone manual sequence improvement and they might contain consensus errors (i.e. substitutions, insertions, and deletions) that impact the annotation of protein-coding genes. For example, the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) has previously sequenced eight *Drosophila* species using the 454 and Illumina sequencing platforms. The 454 sequencing technology exhibits low accuracy in long (> 6bp) mononucleotide runs. Similarly, sequencing reads produced by the Pacific Biosciences (PacBio) and Nanopore sequencing platforms have lower read accuracy (~90%) and higher rates of insertions and deletions compared to Sanger sequencing data.

The GEP developed the *Sequence Updater* tool to provide a standardized way to document consensus errors in the project sequence using the Variant Call Format (VCF). The VCF files generated by the *Sequence Updater* can be used in conjunction with the *Gene Model Checker* to verify gene models that contain consensus sequence errors. Multiple VCF files can be combined into a single project VCF file using the *Annotation Files Merger* in preparation for project submission.

This user guide provides an overview of the strategies that could be used to identify consensus errors in the project sequence, and it describes how to document these consensus errors using the *Sequence Updater*. The *Gene Model Checker* User Guide describes how the VCF file produced by the *Sequence Updater* can be used to verify a gene model with errors in the consensus sequence.

Acknowledgements

The *Sequence Updater* is developed by Wilson Leung at Washington University in St. Louis for the Genomics Education Partnership (GEP).

Questions about the *Sequence Updater*

Please contact Wilson (wleung@wustl.edu) if you have any questions or encounter any problems with the *Sequence Updater*.

Availability

The *Sequence Updater* is available under the “Resources & Tools” section of the F Element project page and the Pathways project page on the GEP website.

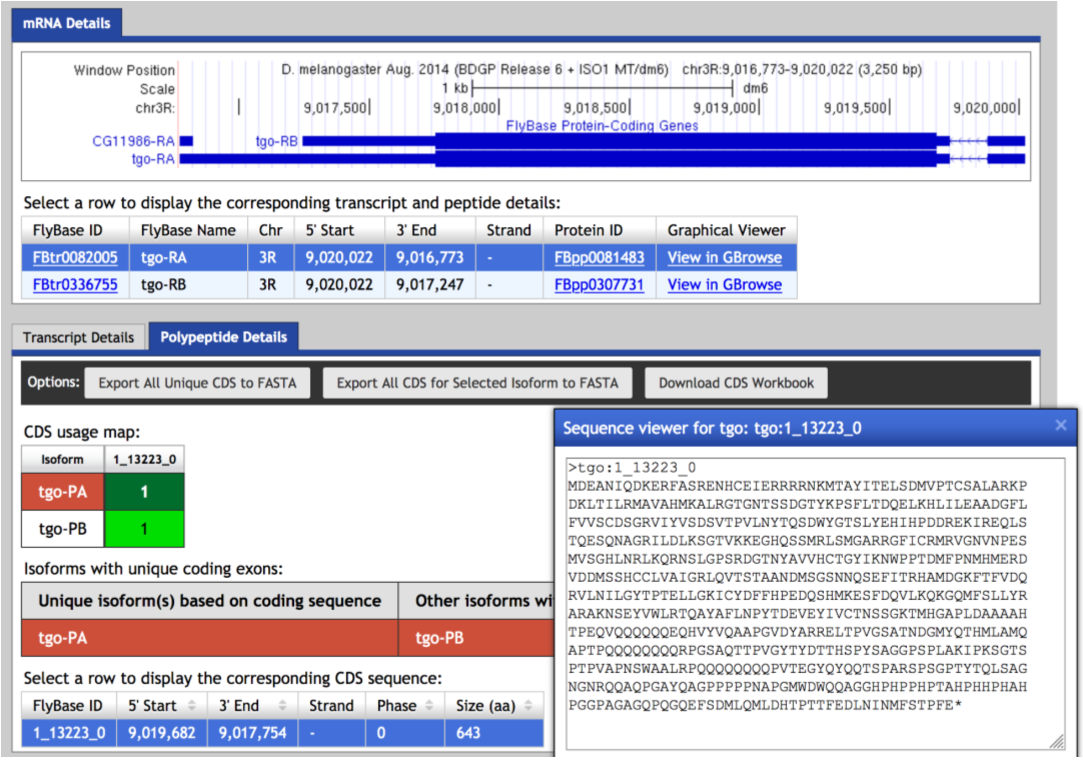
Overview of the *Sequence Updater*

We will use a consensus error within the *tgo* gene in the *Drosophila sechellia* May 2011 (Broad dsec_caf1/DsecCAF1) assembly to illustrate the key functionalities of the *Sequence Updater*. This document will discuss how you can use the *D. sechellia* RNA-Seq data to identify potential discrepancies between the RNA-Seq reads and the project consensus sequence, how to document these errors using the *Sequence Updater*.

Using *tblastn* searches to identify potential consensus sequence errors

When nucleotide insertions or deletions (indels) that are not multiples of three are introduced into the coding regions of a project sequence, the *blastx* or *tblastn* alignment of the coding exons (CDS) against the project sequence will often be split into multiple alignment blocks. This occurs because the indel will introduce a frame-shift within the CDS, which results in the nucleotides before the indel being translated in one reading frame, and the nucleotides after the indel being translated in a different reading frame. Hence the split alignment blocks are typically located adjacent to each other in the project sequence but they are in different reading frames.

We will use the annotation of the *tgo* gene in *D. sechellia* to illustrate the issue with split alignments. The putative ortholog of the *tgo* gene in *D. sechellia* is located at approximately 17,034,485-17,036,408 on scaffold super_0 (GenBank accession number CH480815.1). According to the *Gene Record Finder*, the *tgo* gene has a single CDS (1_13223_0) in *D. melanogaster* (Figure 1).


mRNA Details
Window Position
Scale
chr3R:
D. melanogaster Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) chr3R:9,016,773-9,020,022 (3,250 bp)
1 kb
dm6
9,017,500
9,018,000
9,018,500
9,019,000
9,019,500
9,020,000
FlyBase Protein-Coding Genes
CG11986-RA
tgo-RB
tgo-RA
Transcript Details
Polypeptide Details
Options:
Export All Unique CDS to FASTA
Export All CDS for Selected Isoform to FASTA
Download CDS Workbook

Select a row to display the corresponding transcript and peptide details:

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0082005	tgo-RA	3R	9,020,022	9,016,773	-	FBpp0081483	View in GBrowse
FBtr0336755	tgo-RB	3R	9,020,022	9,017,247	-	FBpp0307731	View in GBrowse

CDS usage map:

Isoform	1_13223_0
tgo-PA	1
tgo-PB	1

Isoforms with unique coding exons:

Unique isoform(s) based on coding sequence	Other isoforms wi
tgo-PA	tgo-PB

Select a row to display the corresponding CDS sequence:

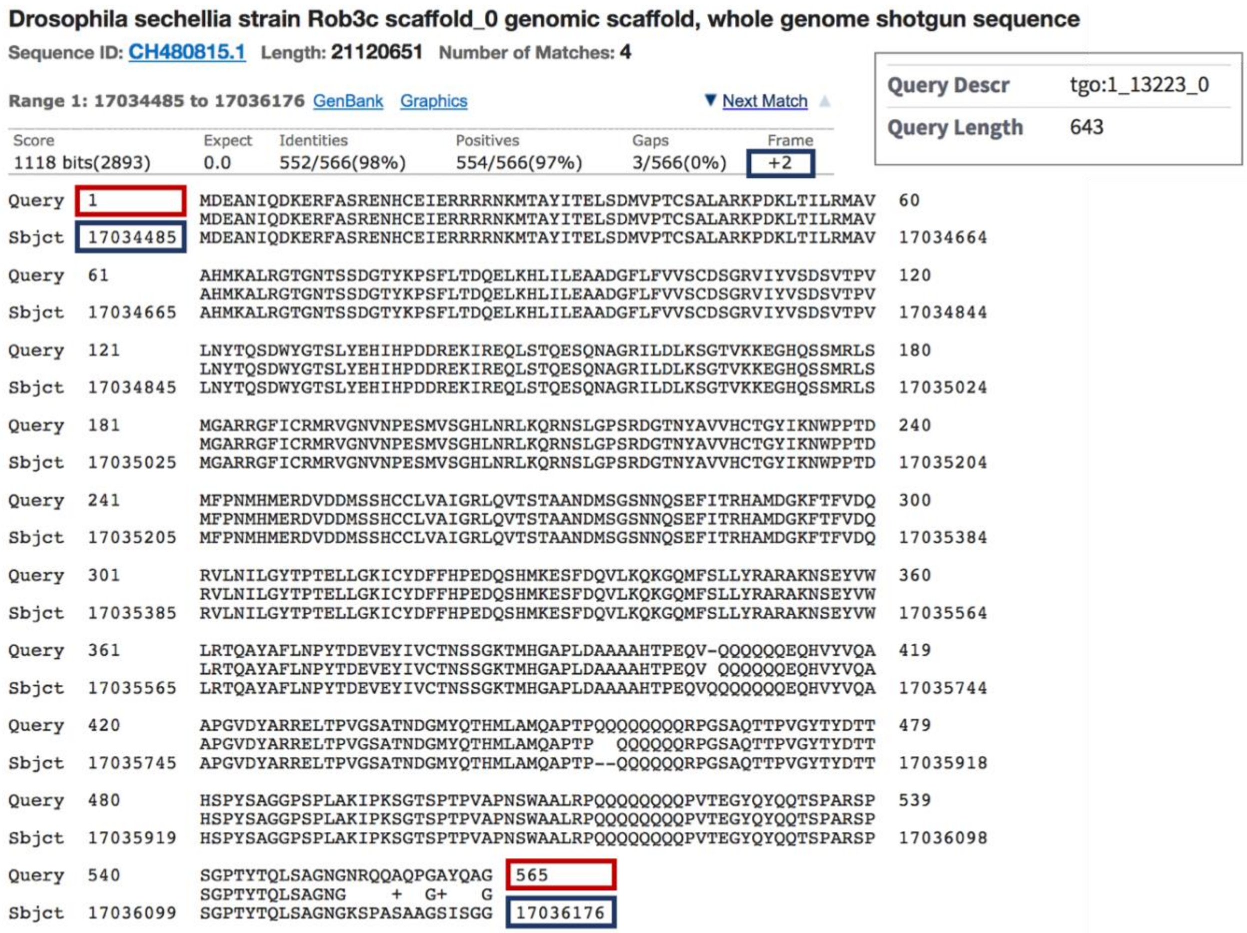
FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_13223_0	9,019,682	9,017,754	-	0	643

Sequence viewer for tgo: tgo:1_13223_0

```
>tgo:1_13223_0
MDEANIQDKERFASRENHCEIERRRRNKMTAYITELSDMVPTCSALARKP
DKLTILRMAVAHMKALRGTGNTSSDGTYKPSFLTDQELKHLILEAADGFL
FVVSCDSGRVIYVSDSVTPVLNYTQSDWYGTSLYEHIHPDDREKIREQLS
TQESQNAGRILDLKSGTVKKEGHQSSMRLSMGARRGFICRMRVGNVNPES
MVSGHLNRLKQRNSLGPSRDGTNYAVVHCTGYIKNWPPTDMFPNMHMERD
VDDMSSHCCLVAIGRLQVTSTAANDMSGSNNQSEFITRHAMDGKFTFVDQ
RVLNILGYTPTELLGKICYDFFHPEDQSHMKESFDQVLKQKGQMFSLLYR
ARAKNSEYVWLRTQAYAFLNPYTDEVEYIVCTNSSGKTMHGAPLDAAAAH
TPEQVQQQQQQEQHVYVQAAPGVDYARRELTPVGSATNDGMYQTHMLAMQ
APTPQQQQQQQQRPGSAQTTPVGYTYDTTHSPYSAGGPSPLAKIPKSGTS
PTPVAPNSWAALRPQQQQQQQQPVTEGYQYQQTSPARSPSGPTYTQLSAG
NGNRQQAQPGAYQAGPPPPPNAPGMWDWQQAGGHPHPPHPTAHPHHPHAH
PGGPAGAGQPQGQEFSDMLQMLDHTPTTFEDLNINMFSTPFE*
```

Figure 1. The *Gene Record Finder* record shows that there are two isoforms of *tgo* in *D. melanogaster* (A and B). Both isoforms have only one CDS (1_13223_0).

A *tblastn* alignment of CDS 1_13223_0 from the *D. melanogaster tgo* gene against the 17,000,000-17,100,000 region of the *D. sechellia* scaffold super_0 shows two highly significant alignment blocks. The first alignment block (with an E-value of 0.0) covers the first 565 residues of CDS 1_13223_0, which spans from 17,034,485-17,036,176 on the *D. sechellia* scaffold super_0 in frame +2 (Figure 2).



Drosophila sechellia strain Rob3c scaffold_0 genomic scaffold, whole genome shotgun sequence

Sequence ID: CH480815.1 Length: 21120651 Number of Matches: 4

Query Descr	tgo:1_13223_0
Query Length	643

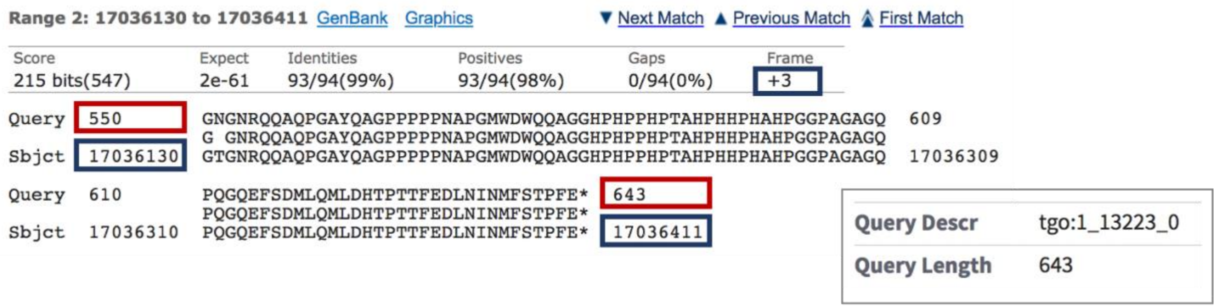
Range 1: 17034485 to 17036176 GenBank Graphics ▼ Next Match

Score	Expect	Identities	Positives	Gaps	Frame
1118 bits(2893)	0.0	552/566(98%)	554/566(97%)	3/566(0%)	+2

```
Query  1         MDEANIQDKERFASRENHCEIERRRRNKMTAYITELSDMVPTCSALARKPDKLTILRMAV  60
                 MDEANIQDKERFASRENHCEIERRRRNKMTAYITELSDMVPTCSALARKPDKLTILRMAV
Sbjct  17034485  MDEANIQDKERFASRENHCEIERRRRNKMTAYITELSDMVPTCSALARKPDKLTILRMAV  17034664

Query  61        AHMKALRGTGNTSSDGTYKPSFLTDQELKHLILEAADGFLFVVSCDSGRVIYVSDSVTPV  120
                 AHMKALRGTGNTSSDGTYKPSFLTDQELKHLILEAADGFLFVVSCDSGRVIYVSDSVTPV
Sbjct  17034665  AHMKALRGTGNTSSDGTYKPSFLTDQELKHLILEAADGFLFVVSCDSGRVIYVSDSVTPV  17034844

Query  121       LNYTQSDWYGTSLYEHIHPDDREKIREQLSTQESQNAGRILDLKSGTVKKEGHQSSMRLS  180
                 LNYTQSDWYGTSLYEHIHPDDREKIREQLSTQESQNAGRILDLKSGTVKKEGHQSSMRLS
Sbjct  17034845  LNYTQSDWYGTSLYEHIHPDDREKIREQLSTQESQNAGRILDLKSGTVKKEGHQSSMRLS  17035024

Query  181       MGARRGFICRMRVGNVNPESMVSGHLNRLKQRNSLGPSRDGTNYAVVHCTGYIKNWPPTD  240
                 MGARRGFICRMRVGNVNPESMVSGHLNRLKQRNSLGPSRDGTNYAVVHCTGYIKNWPPTD
Sbjct  17035025  MGARRGFICRMRVGNVNPESMVSGHLNRLKQRNSLGPSRDGTNYAVVHCTGYIKNWPPTD  17035204

Query  241       MFPNMHMERDVDDMSSHCCLVAIGRLQVTSTAANDMSGSNNQSEFITRHAMDGKFTFVDQ  300
                 MFPNMHMERDVDDMSSHCCLVAIGRLQVTSTAANDMSGSNNQSEFITRHAMDGKFTFVDQ
Sbjct  17035205  MFPNMHMERDVDDMSSHCCLVAIGRLQVTSTAANDMSGSNNQSEFITRHAMDGKFTFVDQ  17035384

Query  301       RVLNILGYTPTELLGKICYDFFHPEDQSHMKESFDQVLKQKGQMFSLLYRARAKNSEYVW  360
                 RVLNILGYTPTELLGKICYDFFHPEDQSHMKESFDQVLKQKGQMFSLLYRARAKNSEYVW
Sbjct  17035385  RVLNILGYTPTELLGKICYDFFHPEDQSHMKESFDQVLKQKGQMFSLLYRARAKNSEYVW  17035564

Query  361       LRTQAYAFLNPYTDEVEYIVCTNSSGKTMHGAPLDAAAAHTPEQV-QQQQQQEQHVYVQA  419
                 LRTQAYAFLNPYTDEVEYIVCTNSSGKTMHGAPLDAAAAHTPEQV QQQQQQEQHVYVQA
Sbjct  17035565  LRTQAYAFLNPYTDEVEYIVCTNSSGKTMHGAPLDAAAAHTPEQVQQQQQQQEQHVYVQA  17035744

Query  420       APGVDYARRELTPVGSATNDGMYQTHMLAMQAPTPQQQQQQQQRPGSAQTTPVGYTYDTT  479
                 APGVDYARRELTPVGSATNDGMYQTHMLAMQAPTP  QQQQQQRPGSAQTTPVGYTYDTT
Sbjct  17035745  APGVDYARRELTPVGSATNDGMYQTHMLAMQAPTP--QQQQQQRPGSAQTTPVGYTYDTT  17035918

Query  480       HSPYSAGGPSPLAKIPKSGTSPTPVAPNSWAALRPQQQQQQQQPVTEGYQYQQTSPARSP  539
                 HSPYSAGGPSPLAKIPKSGTSPTPVAPNSWAALRPQQQQQQQQPVTEGYQYQQTSPARSP
Sbjct  17035919  HSPYSAGGPSPLAKIPKSGTSPTPVAPNSWAALRPQQQQQQQQPVTEGYQYQQTSPARSP  17036098

Query  540       SGPTYTQLSAGNGNRQQAQPGAYQAG  565
                 SGPTYTQLSAGNG    +  G+   G
Sbjct  17036099  SGPTYTQLSAGNGKSPASAAGSISGG  17036176
```

Figure 2. The first alignment block from the *tblastn* search of the *D. melanogaster tgo* CDS 1_13223_0 (query) against the *D. sechellia* scaffold super_0 (CH480815.1; subject) placed the beginning of the CDS at 17,034,485-17,036,176 in frame +2.

The second alignment block covers the residues 550-643 of CDS 1_13223_0, which spans from 17,036,130-17,036,411 on the *D. sechellia* scaffold super_0 in frame +3 (Figure 3).



Range 2: 17036130 to 17036411 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Positives	Gaps	Frame
215 bits(547)	2e-61	93/94(99%)	93/94(98%)	0/94(0%)	+3

```
Query  550       GNGNRQQAQPGAYQAGPPPPPNAPGMWDWQQAGGHPHPPHPTAHPHHPHAHPGGPAGAGQ  609
                 G GNRQQAQPGAYQAGPPPPPNAPGMWDWQQAGGHPHPPHPTAHPHHPHAHPGGPAGAGQ
Sbjct  17036130  GTGNRQQAQPGAYQAGPPPPPNAPGMWDWQQAGGHPHPPHPTAHPHHPHAHPGGPAGAGQ  17036309

Query  610       PQGQEFSDMLQMLDHTPTTFEDLNINMFSTPFE*  643
                 PQGQEFSDMLQMLDHTPTTFEDLNINMFSTPFE*
Sbjct  17036310  PQGQEFSDMLQMLDHTPTTFEDLNINMFSTPFE*  17036411
```

Query Descr	tgo:1_13223_0
Query Length	643

Figure 3. The second alignment block from the *tblastn* search of the *D. melanogaster tgo* CDS 1_13223_0 (query) against the *D. sechellia* scaffold super_0 (CH480815.1; subject) placed the end of the CDS at 17,036,130-17,036,411 in frame +3.

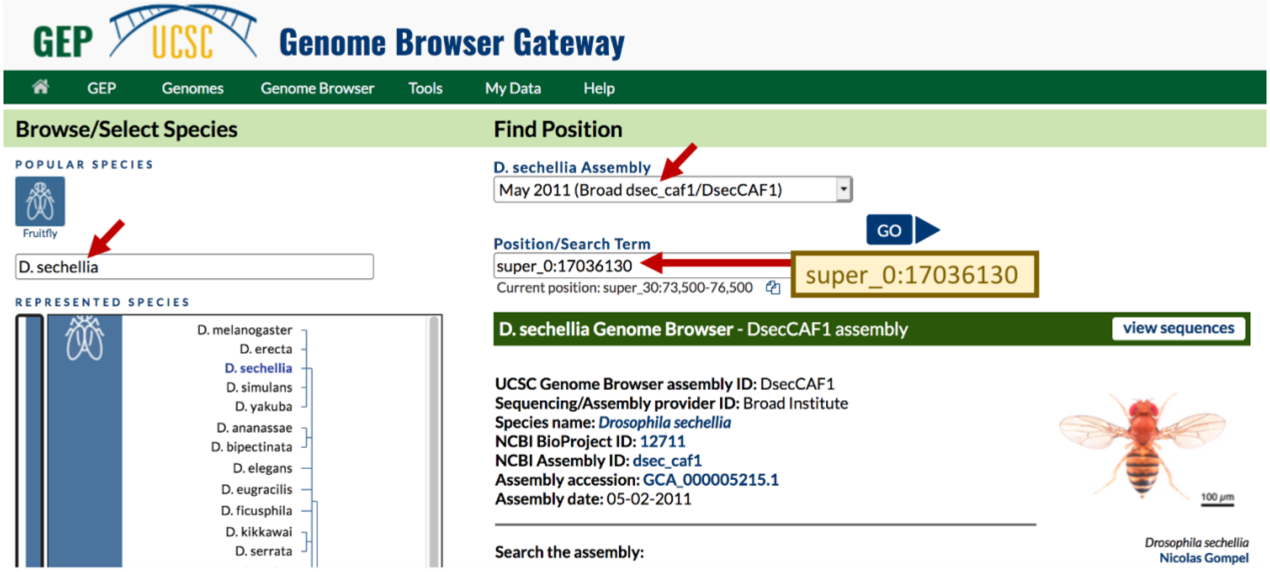
These two alignment blocks cover the entire length (643aa) of CDS 1_13223_0. However, 16 residues (from 550-565) of CDS 1_13223_0 appear in both alignment blocks. In addition, since the first alignment block ends at 17,036,176 and the second alignment block begins at 17,036,130 on the *D. sechellia* scaffold super_0, there is a 47bp overlap between the two alignment blocks. The reading frame has also changed from +2 in the first alignment block to +3 in the second alignment block.

Collectively, the available evidence suggests that CDS 1_13223_0 of the *tgo* gene in *D. sechellia* has either acquired a novel intron or there is an error in the super_0 consensus sequence which leads to a frame shift in the region around 17,036,130 (i.e. the start of the second alignment block).

Using RNA-Seq data to identify potential consensus errors

To determine if the frame shift was caused by a genuine difference in the *D. sechellia* super_0 genomic sequence or an error in the consensus sequence, we will examine the RNA-Seq reads that have been mapped to the region surrounding position 17,036,130 in scaffold super_0. The RNA-Seq reads were generated by the Illumina sequencing platform, which has an average base accuracy greater than 99%. If a genomic region shows multiple high quality discrepancies between the RNA-Seq reads that aligned to the region and the consensus sequence, then either the RNA-Seq reads have been placed in the wrong part of the assembly or the consensus sequence is incorrect.

Based on the start of the second alignment block from the *tblastn* search, the potential issue with the consensus sequence likely occurs at approximately 17,036,130. Open a new web browser window and navigate to the GEP UCSC Genome Browser. Click on the "Genome Browser" link on the left sidebar. Select "*D. sechellia*" under the "Browse/Select Species" section, select "May 2011 (Broad dsec_caf1/DsecCAF1)" under the "D. sechellia Assembly" field, and then enter "super_0:17036130" into the "Position/Search Term" field (Figure 4). Click on the "GO" button.



D. sechellia

D. sechellia Assembly: May 2011 (Broad dsec_caf1/DsecCAF1)

Position/Search Term: super_0:17036130

Current position: super_30:73,500-76,500

D. sechellia Genome Browser - DsecCAF1 assembly

UCSC Genome Browser assembly ID: DsecCAF1
Sequencing/Assembly provider ID: Broad Institute
Species name: *Drosophila sechellia*
NCBI BioProject ID: 12711
NCBI Assembly ID: dsec_caf1
Assembly accession: GCA_000005215.1
Assembly date: 05-02-2011

Search the assembly:

Figure 4. Navigate to position super_0:17036130 in the *D. sechellia* May 2011 (Broad dsec_caf1/DsecCAF1) assembly using the *GEP UCSC Genome Browser*.

To get a broader view of the genomic region surrounding this position, zoom out 100x. Click on the "hide all" button beneath the Genome Browser image to hide all the evidence tracks. Scroll down to the track configuration section and then change the display settings for the following evidence tracks:

Under "Mapping and Sequencing Tracks":

- Base Position: **dense**
- Gap: **dense**

Under "Genes and Gene Prediction Tracks":

- RefSeq Genes: **pack**
- D. mel Proteins: **pack**
- *GeMoMa* Genes: **pack**
- *Augustus*: **pack**

Under "RNA-Seq Tracks":

- RNA-Seq Coverage: **full**

Click on one of the "refresh" buttons to update the Genome Browser display.

The *GeMoMa* prediction FBTR0082005_R0 and the *Augustus* gene prediction super_0.g1981.t1 both introduced an intron in this region. The *BLAT* alignment of XM_032720520 in the NCBI "RefSeq Genes" track shows a single base gap. The "RNA-Seq Coverage" tracks show a substantial drop in read coverage in the adult females and adult males RNA-Seq samples at the same position as the gap in the "RefSeq Genes" track (Figure 5).

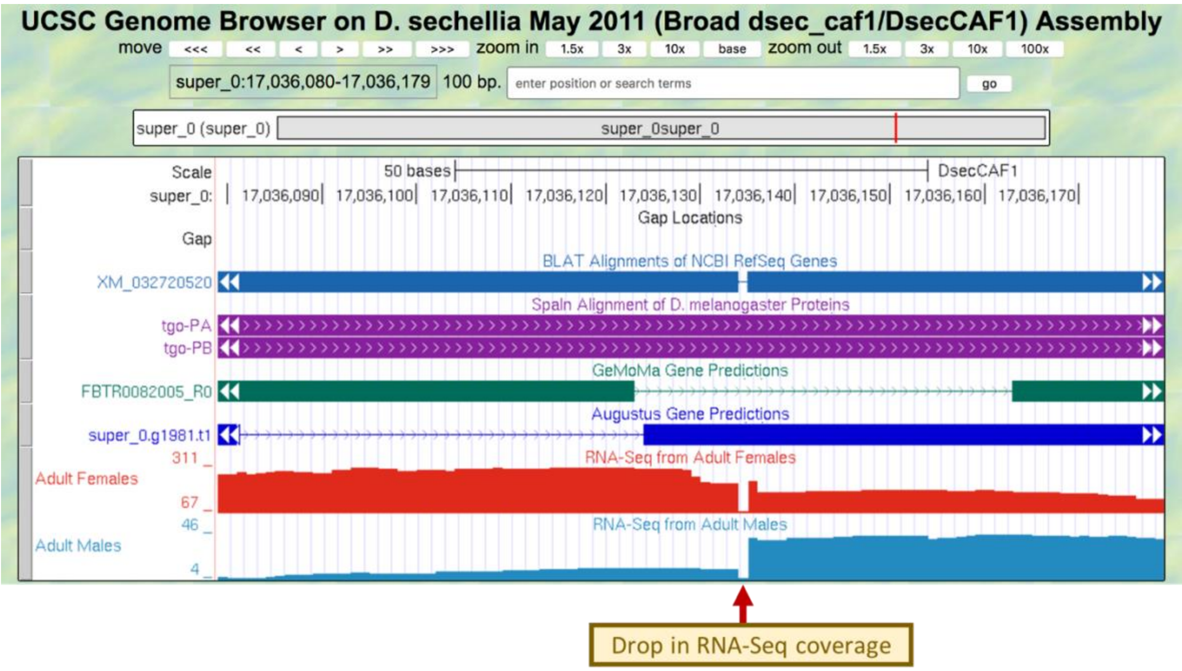

UCSC Genome Browser on D. sechellia May 2011 (Broad dsec_caf1/DsecCAF1) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
super_0:17,036,080-17,036,179 100 bp.
enter position or search terms
go
super_0 (super_0)
super_0super_0
Scale
50 bases
DsecCAF1
super_0:
17,036,090
17,036,100
17,036,110
17,036,120
17,036,130
17,036,140
17,036,150
17,036,160
17,036,170
Gap Locations
Gap
BLAT Alignments of NCBI RefSeq Genes
XM_032720520
Spaln Alignment of D. melanogaster Proteins
tgo-PA
tgo-PB
GeMoMa Gene Predictions
FBTR0082005_R0
Augustus Gene Predictions
super_0.g1981.t1
311
RNA-Seq from Adult Females
Adult Females
67
46
RNA-Seq from Adult Males
Adult Males
4
Drop in RNA-Seq coverage

Figure 5. Introns predicted by multiple gene predictors and the decrease in RNA-Seq read coverage demarcate the position within the *D. sechellia* super_0 scaffold which contains the potential consensus error.

To inspect the position which shows a substantial decrease in RNA-Seq coverage more closely, enter "super_0:17,036,125-17,036,145" into the "enter position or search terms" text box and then click on the "go" button. We find that G nucleotide at 17,036,135 corresponds to the position with the substantial drop in RNA-Seq read coverage (Figure 6). This drop in RNA-Seq read coverage typically indicates a discrepancy between the RNA-Seq reads aligned to this region and the consensus sequence.

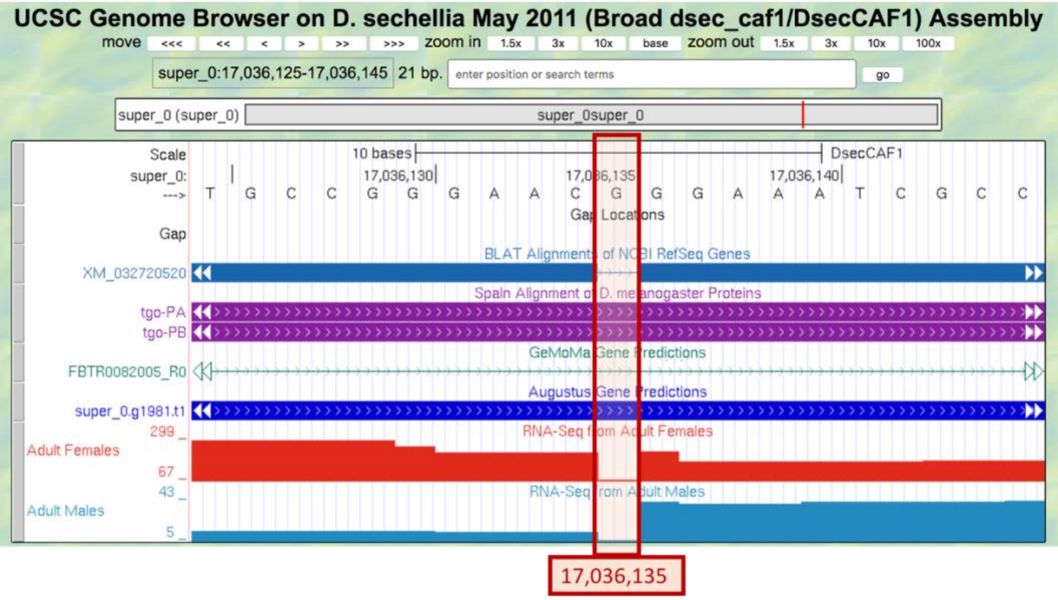
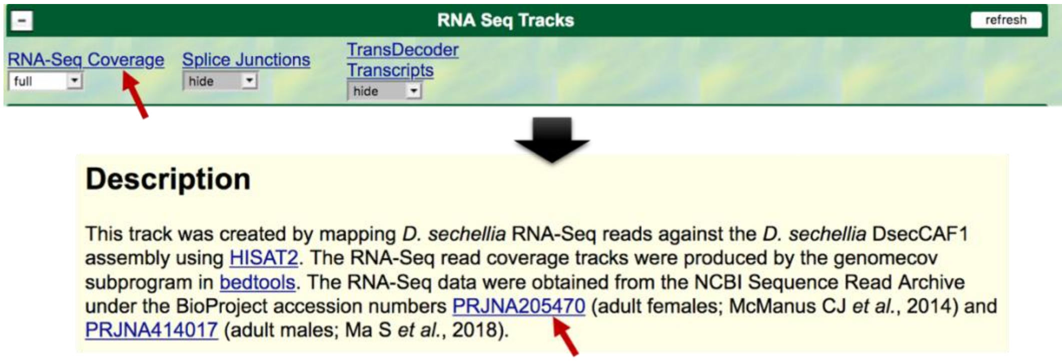

UCSC Genome Browser on D. sechellia May 2011 (Broad dsec_caf1/DsecCAF1) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
super_0:17,036,125-17,036,145 21 bp. enter position or search terms go
super_0 (super_0) super_0super_0
Scale
super_0:
--->
10 bases
DsecCAF1
17,036,130
17,036,135
17,036,140
T G C C G G G A A C G G G A A A T C G C C
Gap Locations
Gap
BLAT Alignments of NCBI RefSeq Genes
XM_032720520
Spaln Alignment of D. melanogaster Proteins
tgo-PA
tgo-PB
GeMoMa Gene Predictions
FBTR0082005_R0
Augustus Gene Predictions
super_0.g1981.t1
RNA-Seq from Adult Females
Adult Females
299
67
RNA-Seq from Adult Males
Adult Males
43
5
17,036,135

Figure 6. The RNA-Seq Read coverage tracks indicate that the discrepant position between the RNA-Seq reads and the consensus sequence is located at 17,036,135 in scaffold super_0.

Use *SRA-BLAST* to gather additional evidence for the consensus error

To assess the potential consensus sequence error at position 17,036,135 in scaffold super_0, we will compare this genomic region against the *D. sechellia* RNA-Seq reads using NCBI *BLAST*.

In order to perform this analysis, we will need to first determine the source of the RNA-Seq data used to generate the RNA-Seq read coverage track. Scroll down to the "RNA-Seq Tracks" section and then click on the "RNA-Seq Coverage" link. The "Description" section indicates that the adult females RNA-Seq data was obtained from the BioProject with the accession number PRJNA205470, and the adult males RNA-Seq data was obtained from the BioProject PRJNA414017.

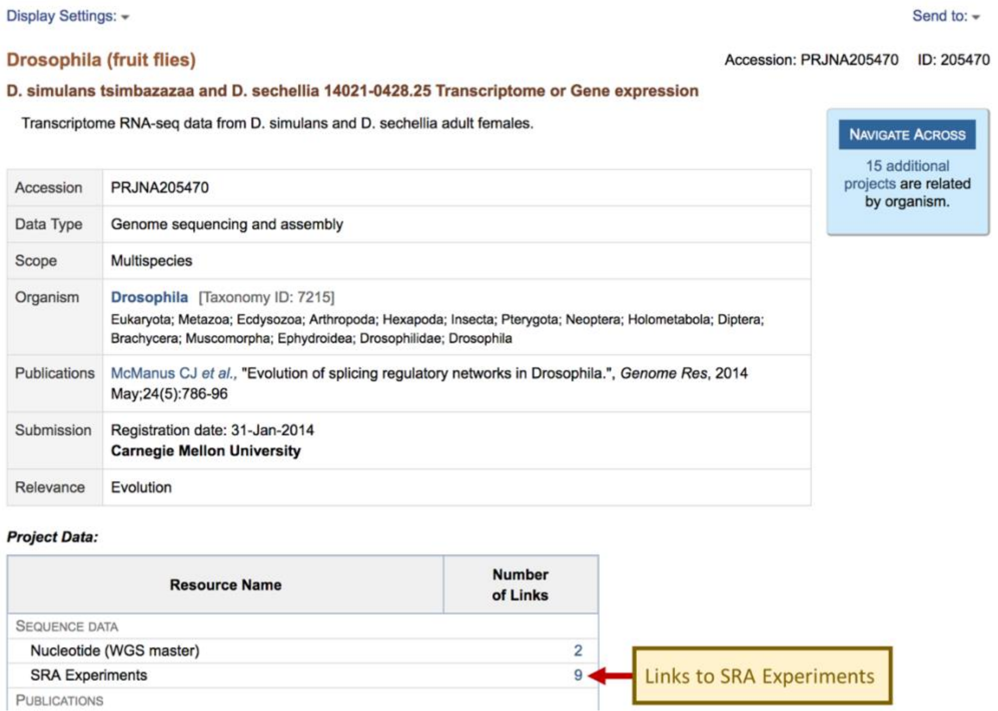


Description

This track was created by mapping *D. sechellia* RNA-Seq reads against the *D. sechellia* DsecCAF1 assembly using HISAT2. The RNA-Seq read coverage tracks were produced by the genomecov subprogram in bedtools. The RNA-Seq data were obtained from the NCBI Sequence Read Archive under the BioProject accession numbers PRJNA205470 (adult females; McManus CJ *et al.*, 2014) and PRJNA414017 (adult males; Ma S *et al.*, 2018).

Figure 7. The "Description" section of the details page for the "RNA-Seq Coverage" track indicates that the *D. sechellia* adult females RNA-Seq data was obtained from the NCBI BioProject PRJNA205470.

Click on the "PRJNA205470" link in the "Description" section to navigate to the BioProject record for the *D. sechellia* RNA-Seq sample. Scroll down to the "SRA Experiment" row under the "Project Data" section, and then click on the "9" link (Figure 8).



Display Settings: ▾ Send to: ▾

Drosophila (fruit flies) Accession: PRJNA205470 ID: 205470

D. simulans tsimbazazaa and D. sechellia 14021-0428.25 Transcriptome or Gene expression

Transcriptome RNA-seq data from D. simulans and D. sechellia adult females.

Accession	PRJNA205470
Data Type	Genome sequencing and assembly
Scope	Multispecies
Organism	**Drosophila** [Taxonomy ID: 7215] Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Holometabola; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila
Publications	McManus CJ *et al.*, "Evolution of splicing regulatory networks in Drosophila.", *Genome Res*, 2014 May;24(5):786-96
Submission	Registration date: 31-Jan-2014 **Carnegie Mellon University**
Relevance	Evolution

NAVIGATE ACROSS

15 additional projects are related by organism.

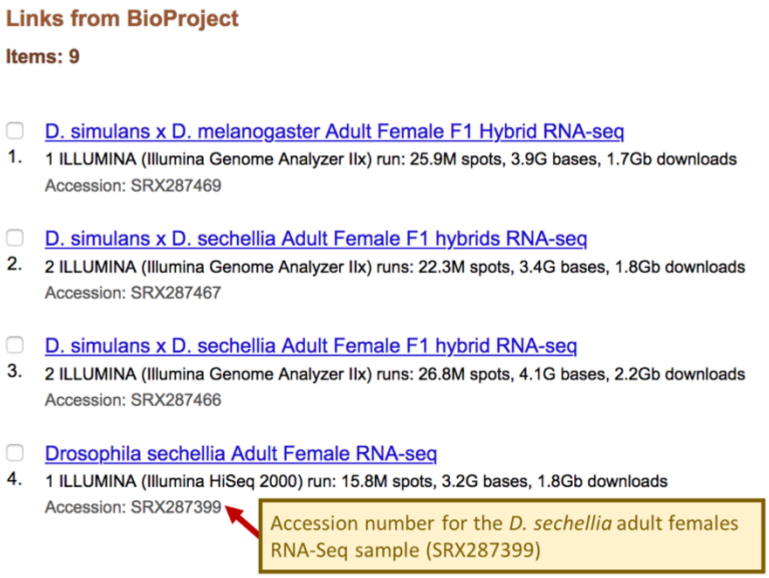
Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (WGS master)	2
SRA Experiments	9
PUBLICATIONS	

Links to SRA Experiments

Figure 8. Click on the link under the "SRA Experiments" row in the "Project Data" table to access the sequencing data associated with the BioProject PRJNA205470.

The NCBI Sequence Read Archive (SRA) search results page shows that the *D. sechellia* adult females RNA-Seq data has the SRA Experiment (SRX) accession number SRX287399 (Figure 9).



Links from BioProject

Items: 9

☐ D. simulans x D. melanogaster Adult Female F1 Hybrid RNA-seq
1. 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 25.9M spots, 3.9G bases, 1.7Gb downloads
Accession: SRX287469

☐ D. simulans x D. sechellia Adult Female F1 hybrids RNA-seq
2. 2 ILLUMINA (Illumina Genome Analyzer IIx) runs: 22.3M spots, 3.4G bases, 1.8Gb downloads
Accession: SRX287467

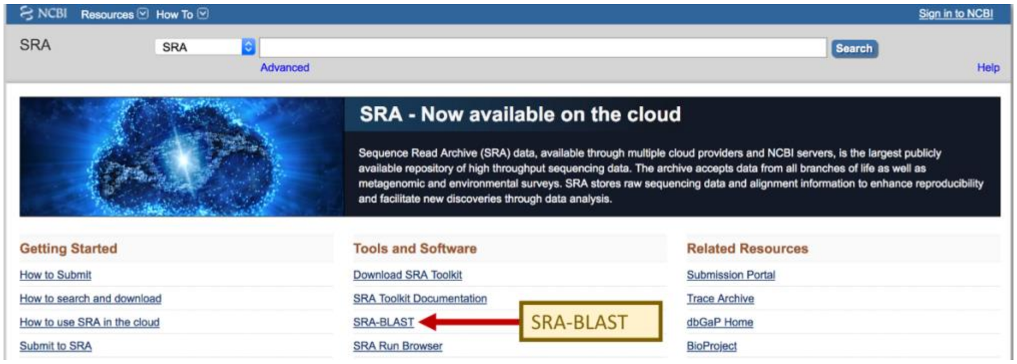
☐ D. simulans x D. sechellia Adult Female F1 hybrid RNA-seq
3. 2 ILLUMINA (Illumina Genome Analyzer IIx) runs: 26.8M spots, 4.1G bases, 2.2Gb downloads
Accession: SRX287466

☐ Drosophila sechellia Adult Female RNA-seq
4. 1 ILLUMINA (Illumina HiSeq 2000) run: 15.8M spots, 3.2G bases, 1.8Gb downloads
Accession: SRX287399

Accession number for the *D. sechellia* adult females RNA-Seq sample (SRX287399)

Figure 9. Use the SRA search results page to determine the accession number for the *D. sechellia* adult females RNA-Seq experiment (i.e. SRX287399).

In order to compare the *D. sechellia* RNA-Seq reads from this experiment against the consensus sequence from the super_0 scaffold, open a new web browser window and navigate to the NCBI SRA home page. Click on the "SRA-BLAST" link under the "Tools and Software" section (Figure 10).



SRA - Now available on the cloud

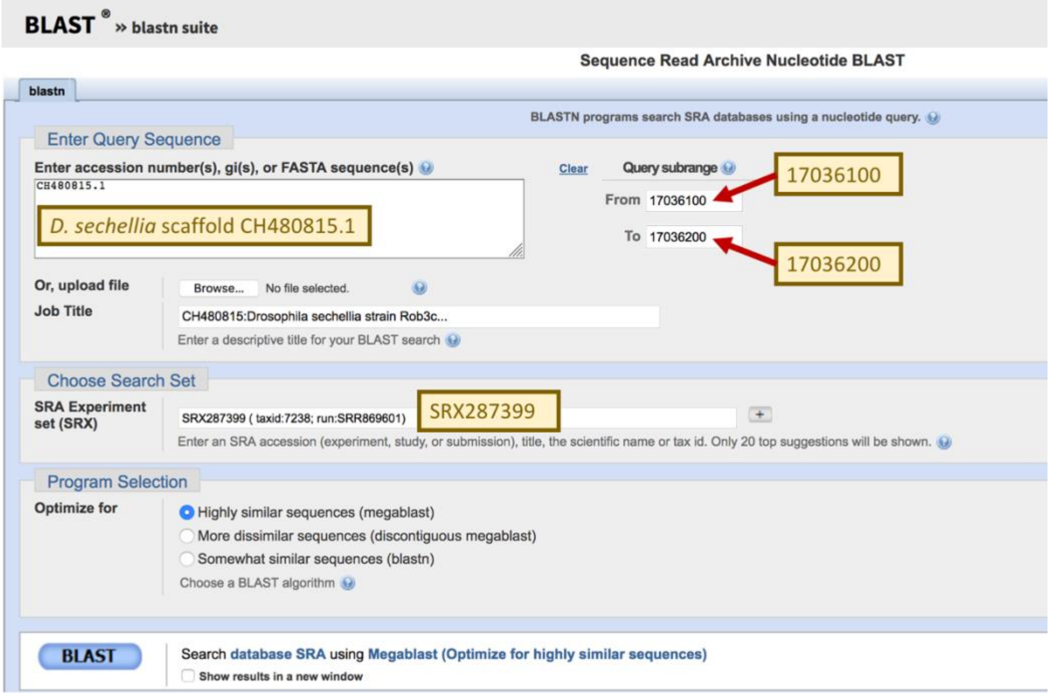
Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started	Tools and Software	Related Resources
How to Submit	Download SRA Toolkit	Submission Portal
How to search and download	SRA Toolkit Documentation	Trace Archive
How to use SRA in the cloud	SRA-BLAST	dbGaP Home
Submit to SRA	SRA Run Browser	BioProject

Figure 10. Access the *SRA-BLAST* service from the NCBI SRA home page.

Enter the GenBank accession number for the *D. sechellia* scaffold super_0 (i.e. CH480815.1) into the "Enter Query Sequence" text box. Since the potential discrepant position is located at 17,036,135, we will limit the search region to 17,036,100-17,036,200 of the scaffold. Under the "Query subrange" section, enter "17036100" into the "From" field and "17036200" into the "To" field.

Under the "SRA Experiment set (SRX)" field, enter the accession number "SRX287399" to search the reads from the *D. sechellia* adult females RNA-Seq experiment. As you enter the accession number, the field will show a suggestion with the taxonomy and SRA Run accession number for the experiment [i.e. SRX287399 (taxid:7238; run:SRR869601)] (Figure 11). Click on the "BLAST" button to run the search.


BLAST ® » blastn suite
Sequence Read Archive Nucleotide BLAST
blastn
BLASTN programs search SRA databases using a nucleotide query.
Enter Query Sequence
Clear
Browse...
Choose Search Set
Program Selection
BLAST
D. sechellia scaffold CH480815.1
17036100
17036200
SRX287399

Enter accession number(s), gi(s), or FASTA sequence(s): CH480815.1

Query subrange
From: 17036100
To: 17036200

Or, upload file: No file selected.

Job Title: CH480815:Drosophila sechellia strain Rob3c...
Enter a descriptive title for your BLAST search

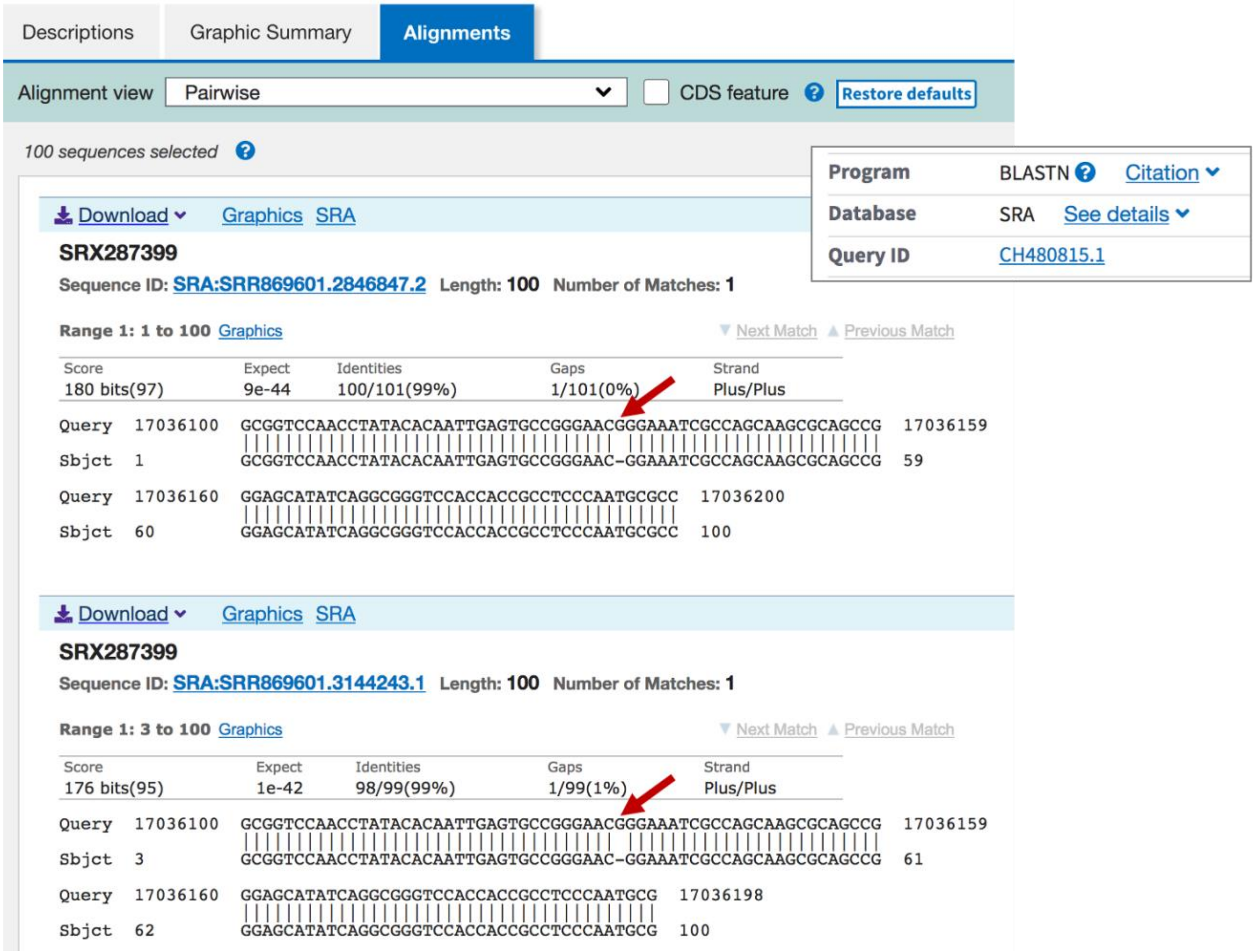
SRA Experiment set (SRX): SRX287399 (taxid:7238; run:SRR869601)
Enter an SRA accession (experiment, study, or submission), title, the scientific name or tax id. Only 20 top suggestions will be shown.

Optimize for
● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)
Choose a BLAST algorithm

Search database SRA using Megablast (Optimize for highly similar sequences)
□ Show results in a new window

Figure 11. Configure *SRA-BLAST* to search the 17,036,100-17,036,200 region of the *D. sechellia* scaffold super_0 (CH480815.1; query) against the sequencing reads in the SRA Experiment SRX287399 (database).

Once the *SRA-BLAST* search is complete, click on the “Alignments” tab to see the alignments between the *D. sechellia* RNA-Seq reads and the consensus sequence. These alignments consistently show an extra G at 17,036,135 of the *D. sechellia* scaffold super_0 compared to the RNA-Seq reads (Figure 12).



Alignment view: Pairwise ☐ CDS feature

100 sequences selected

Program	BLASTN	Citation
Database	SRA	See details
Query ID	CH480815.1	

SRX287399

Sequence ID: SRA:SRR869601.2846847.2 Length: 100 Number of Matches: 1

Range 1: 1 to 100 Graphics

Score	Expect	Identities	Gaps	Strand
180 bits(97)	9e-44	100/101(99%)	1/101(0%)	Plus/Plus

```
Query  17036100  GCGGTCCAACCTATACACAATTGAGTGCCGGGAACGGGAAATCGCCAGCAAGCGCAGCCG  17036159
                 ||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  1         GCGGTCCAACCTATACACAATTGAGTGCCGGGAAC-GGAAATCGCCAGCAAGCGCAGCCG  59

Query  17036160  GGAGCATATCAGGCGGGTCCACCACCGCCTCCCAATGCGCC  17036200
                 |||||||||||||||||||||||||||||||||||||||||
Sbjct  60        GGAGCATATCAGGCGGGTCCACCACCGCCTCCCAATGCGCC  100
```

SRX287399

Sequence ID: SRA:SRR869601.3144243.1 Length: 100 Number of Matches: 1

Range 1: 3 to 100 Graphics

Score	Expect	Identities	Gaps	Strand
176 bits(95)	1e-42	98/99(99%)	1/99(1%)	Plus/Plus

```
Query  17036100  GCGGTCCAACCTATACACAATTGAGTGCCGGGAACGGGAAATCGCCAGCAAGCGCAGCCG  17036159
                 ||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  3         GCGGTCCAACCTATACACAATTGAGTGCCGGGAAC-GGAAATCGCCAGCAAGCGCAGCCG  61

Query  17036160  GGAGCATATCAGGCGGGTCCACCACCGCCTCCCAATGCG  17036198
                 |||||||||||||||||||||||||||||||||||||||
Sbjct  62        GGAGCATATCAGGCGGGTCCACCACCGCCTCCCAATGCG  100
```

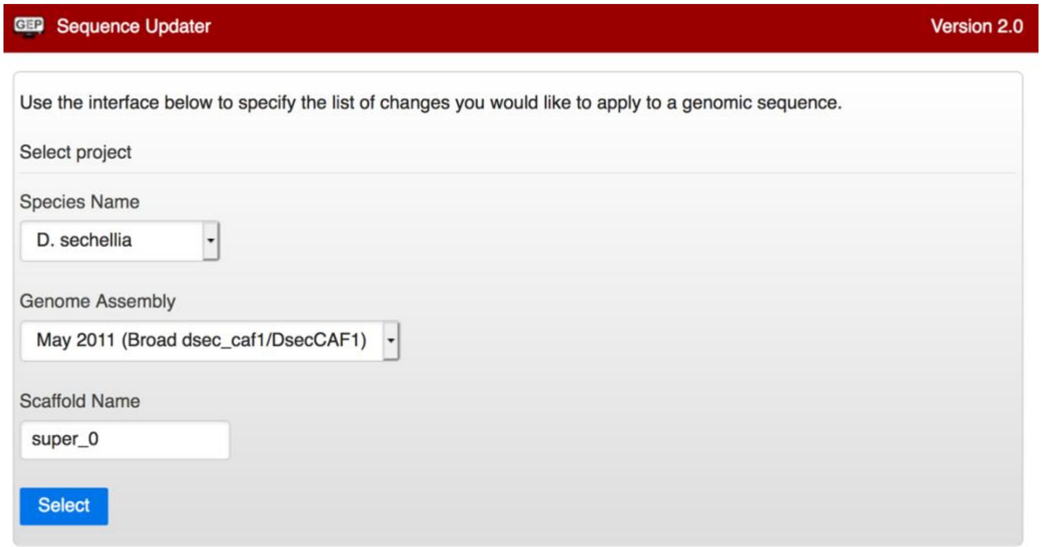
Figure 12. The *SRA-BLAST* alignment of the *D. sechellia* scaffold super_0 (CH480815.1; query) against the RNA-Seq read SRR869601.2846847.2 (top) and SRR869601.3144243.1 (bottom). Both alignments show an extra G at 17,036,135 in the *D. sechellia* scaffold super_0.

Based on the analysis of the *tblastn* search result for CDS 1_13223_0, the gene predictions and RNA-Seq read coverage tracks on the *GEP UCSC Genome Browser*, and the *SRA-BLAST* search result, the G at 17,036,135 should be removed from the *D. sechellia* super_0 sequence.

Document the consensus sequence error with the *Sequence Updater*

Now that we have identified a consensus error at position 17,036,135 of scaffold super_0, we will use the *Sequence Updater* to document this error. Open a new web browser tab and navigate to the Pathways project page on the GEP website. Click on the "*Sequence Updater*" link under the "Resources & Tools" section.

Select "D. sechellia" under the "Species Name" field, select "May 2011 (Broad dsec_caf1/DsecCAF1) under the "Genome Assembly" field, and then enter "super_0" into the "Scaffold Name" field (Figure 13). Click on the "Select" button.



Use the interface below to specify the list of changes you would like to apply to a genomic sequence.

Select project

Species Name: D. sechellia

Genome Assembly: May 2011 (Broad dsec_caf1/DsecCAF1)

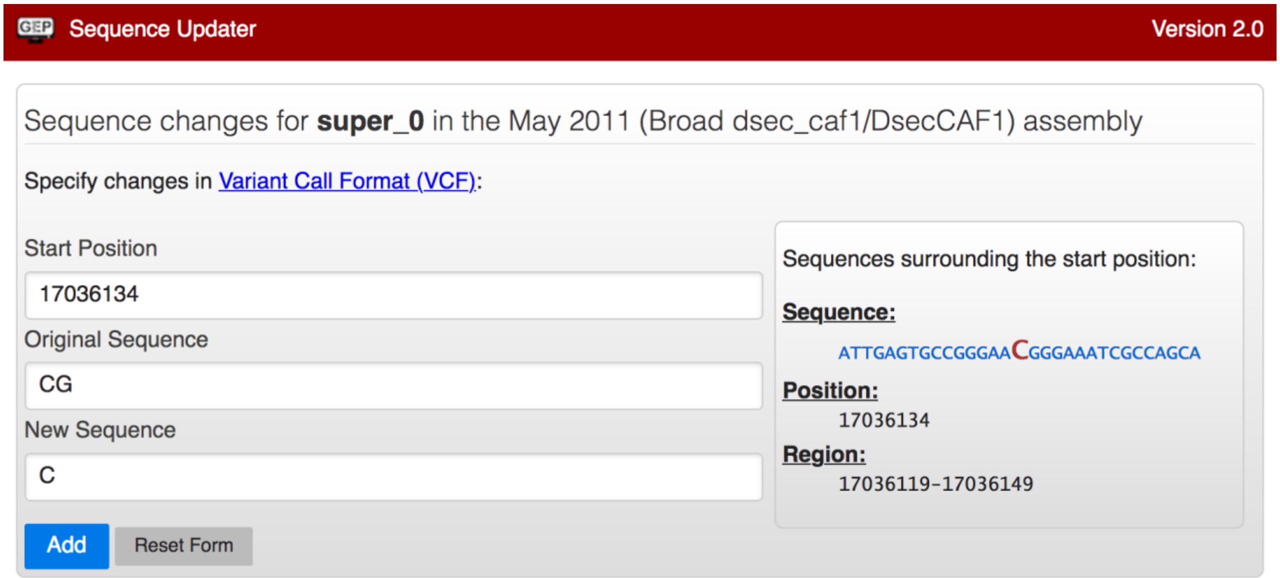
Scaffold Name: super_0

Figure 13. Specify the species, genome assembly, and the scaffold that will be modified by the *Sequence Updater*.

The *Sequence Updater* uses the Variant Call Format (VCF) to describe the changes to the original project sequence. In order to document a change in the sequence, we must specify the start coordinate of the change (relative to the original sequence), the original sequence, and the new sequence. The start coordinate corresponds to the first position that changed between the original and the new sequence. In the case of base substitutions, this will correspond to the position where the first base substitution occurs. However, in the case of base insertions or deletions, the start position will correspond to the base just before the indels.

Since we would like to remove the G nucleotide at 17,036,135 from the consensus sequence, we will use 17,036,134 as the start coordinate, and then omit the G at 17,036,135 as part of the change.

Enter "17036134" into the "Start Position" field. A tooltip will appear which shows the nucleotide at this position (C) and the nucleotides surrounding this position. To remove the G at 17,036,135, enter "CG" into the "Original Sequence" field and "C" into the "New Sequence" field (Figure 14). Click on the "Add" button to add the proposed modification to the "List of sequence changes" section (Figure 15).



Sequence Updater Version 2.0

Sequence changes for **super_0** in the May 2011 (Broad dsec_caf1/DsecCAF1) assembly

Specify changes in Variant Call Format (VCF):

Start Position: 17036134

Original Sequence: CG

New Sequence: C

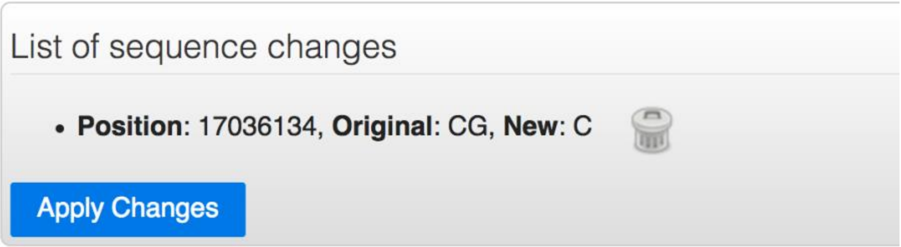
Sequences surrounding the start position:

Sequence:
ATTGAGTGCCGGGAACGGGAAATCGCCAGCA

Position:
17036134

Region:
17036119-17036149

Figure 14. Use the *Sequence Updater* interface to describe the changes that should be applied to the original sequence. When you specify a "Start Position", a tooltip will appear on the right which shows the nucleotide at that position (e.g., the C in red) and the 15 nucleotides before and after that position (nucleotides in blue).



List of sequence changes

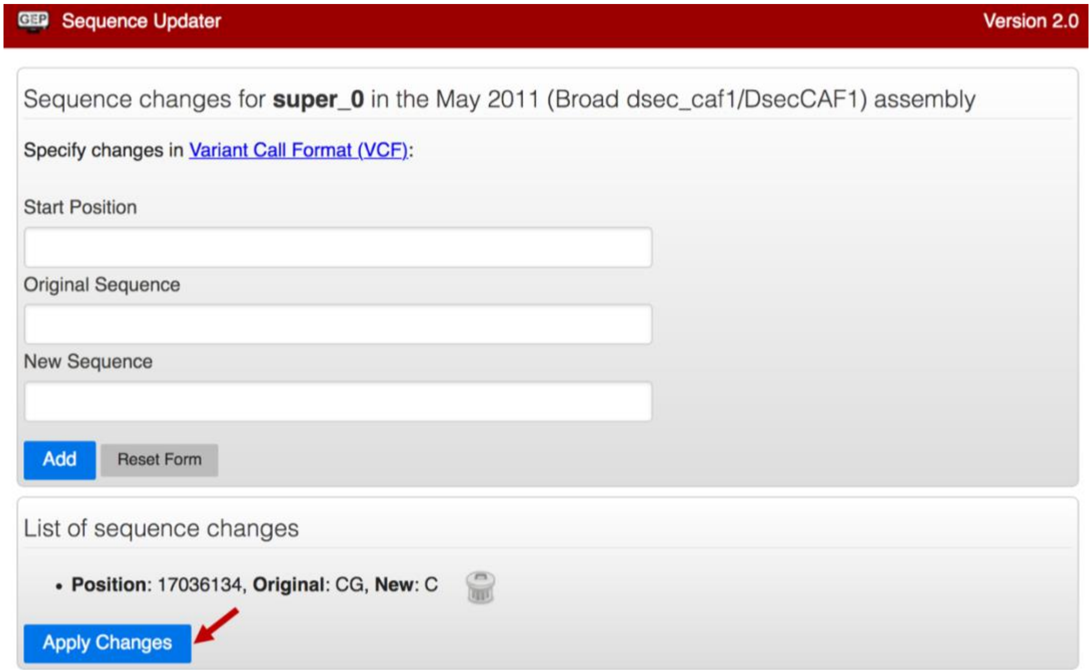
- **Position**: 17036134, **Original**: CG, **New**: C

Figure 15. The changes that will be applied to the original sequence are listed under the "List of sequence changes" section. Click on the trash can icon to delete the modification from the list of sequence changes.

If the region contains multiple consensus errors, you can use the same procedure to add the additional changes to the "List of sequence changes" section.

Note: The "Start Position" and "Original Sequence" are based on the sequence in the *GEP UCSC Genome Browser*. In cases where there are multiple consensus errors, **all the modifications should be relative to the original sequence** in the *GEP UCSC Genome Browser*. The *Sequence Updater* will automatically transforms the provided start positions as it iteratively applies the modifications to the original sequence.

Click on the "Apply Changes" button to create the VCF file which describes the list of changes to the original sequence (Figure 16).



Sequence changes for **super_0** in the May 2011 (Broad dsec_caf1/DsecCAF1) assembly

Specify changes in Variant Call Format (VCF):

Start Position

Original Sequence

New Sequence

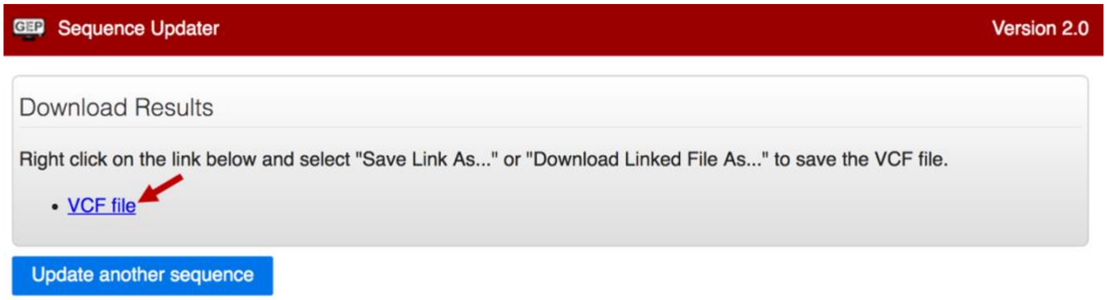
List of sequence changes

- **Position**: 17036134, **Original**: CG, **New**: C

Figure 16. After specifying the sequence changes using the *Sequence Updater* interface, click on the "Apply Changes" button to generate the VCF and the revised sequence file.

Once the analysis is complete, A "Download Results" panel will appear with a link to the VCF file (Figure 17). Right click (control-click on macOS) on the "VCF file" link and then select "Save Link As…" or "Download Linked File As…" to save the VCF file on your computer.

For projects with errors in the consensus sequence, you will need to submit the VCF file in conjunction with the annotation report, the GFF file, the transcript sequence file, and the peptide sequence file. For the purpose of this tutorial, we will name the VCF file "**dsec_tgo.vcf.txt**".



Download Results

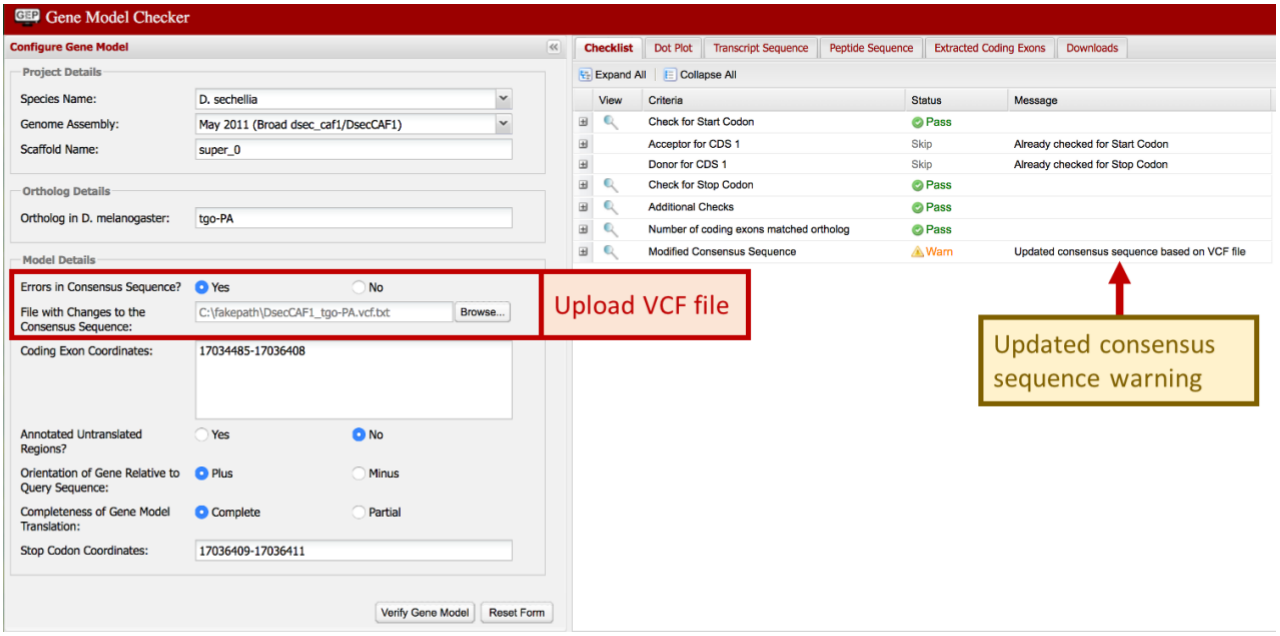
Right click on the link below and select "Save Link As..." or "Download Linked File As..." to save the VCF file.

- VCF file

Figure 17. Right-click (control-click on macOS) on the "VCF file" link and select "Save Link As..." or "Download Linked File As…" to save the VCF file generated by the *Sequence Updater*.

Use the VCF file to verify gene models with errors in the consensus sequence

After documenting the error at 17,036,135 of the *D. sechellia* scaffold super_0, we can apply the changes in the VCF file to the project sequence when we verify a gene model using the *Gene Model Checker*. The *Gene Model Checker* User Guide includes a walkthrough which illustrates how to use the VCF file with the *Gene Model Checker* to verify the gene model for *D. sechellia* tgo-PA (Figure 18).



Project Details

Species Name:	D. sechellia
Genome Assembly:	May 2011 (Broad dsec_caf1/DsecCAF1)
Scaffold Name:	super_0

Ortholog Details

Ortholog in D. melanogaster:	tgo-PA

Model Details

Errors in Consensus Sequence?	● Yes	○ No
File with Changes to the Consensus Sequence:	C:\fakepath\DsecCAF1_tgo-PA.vcf.txt	
Coding Exon Coordinates:	17034485-17036408	
Annotated Untranslated Regions?	○ Yes	● No
Orientation of Gene Relative to Query Sequence:	● Plus	○ Minus
Completeness of Gene Model Translation:	● Complete	○ Partial
Stop Codon Coordinates:	17036409-17036411	

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 ortholog	✓ Pass	
🔍	Modified Consensus Sequence	⚠ Warn	Updated consensus sequence based on VCF file

Figure 18. The VCF file produced by the *Sequence Updater* can be provided to the *Gene Model Checker* through the "File with Changes to the Consensus Sequence" field.

Note: When you use the *Gene Model Checker* to verify a gene model with consensus errors, all the exon coordinates should be **relative to the original project sequence**.

Conclusion

This user guide demonstrates how you can use the *Sequence Updater* to document potential consensus errors in the project sequence. It shows how *tblastn* CDS searches and RNA-Seq data on the *GEP UCSC Genome Browser* can be used to identify potential consensus errors. It then uses the NCBI SRA and the *SRA-BLAST* service to compare the RNA-Seq reads against the project sequence to confirm a consensus error. Finally, the user guide provides an overview of the *Sequence Updater* interface and how to use it to create a VCF file that is suitable for use with the *Gene Model Checker*.