CPGAVAS2 HELP

Version 2.0

Last update: April 15th, 2019

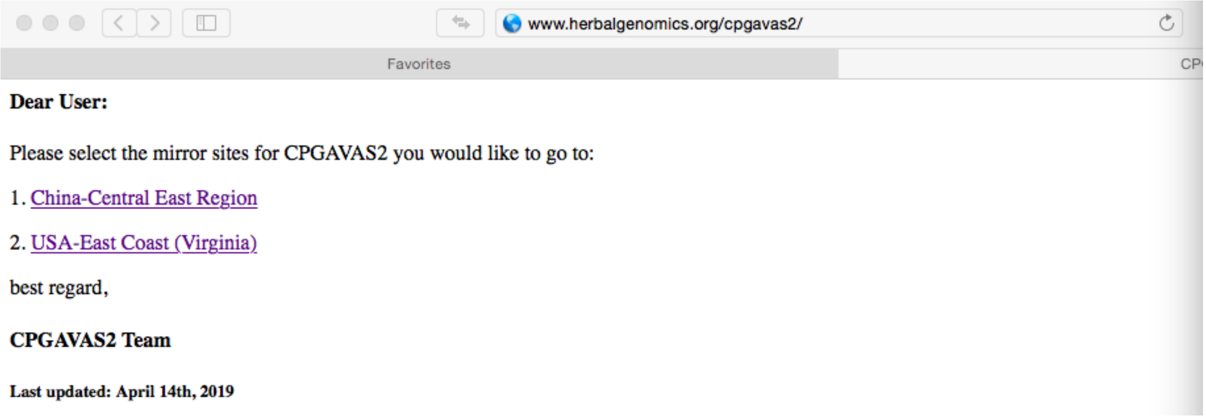
Table of Contents

INTRODUCTION

CPGAVAS2 is a significant improvement of CPGAVAS and contains five main features.

1. "AnnotateGenome"- The user needs to provide a FASTA sequence and optional GenBank file, the plastome annotation will be carried out.
2. "ViewResults"- This module allows retrieval and examination of the annotation results for all of the analysis module including in CPGVAS2.
3. "UpdateAnnotatioinResults"- The manually curated gene annotation information in GFF3 format file can be re-analyzed using this function. It will reproduce the circular map and the analysis results.
4. "ExtractSeq"- With the availability of more than 3000 plastomes, phylogenetic analyses can be used to understand taxonomic relationship between the newly obtained plastome and those having already been sequences.
5. "AnaDiversity"- This module supports the preliminary identification of Single nucleotide polymorphisms (SNP) and the prediction of RNA editing sites using NGS data. However, the results will depend on the setup of particular experiments.

The web server can be accessed at http://www.herbalgenomics.org/cpgavas2. There are two mirror sites for CPGAVAS2, users can select any one according to their network connection.



Dear User:

Please select the mirror sites for CPGAVAS2 you would like to go to:

1. China-Central East Region

2. USA-East Coast (Virginia)

best regard,

CPGAVAS2 Team

Last updated: April 14th, 2019

If you use this webserver, please cite the following paper.

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., & Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC genomics, 13(1), 715.

1 INPUT files for CPGAVAS2

CPGAVAS2 will takes four kinds of input files for genome annotation, repeat analysis, and diversity analysis. Diversity analysis can be divided into two pipelines: SNP discovery and RNA editing site discovery. Below we will explain the files required for these analysis. Samples files are provided near where the user need to upload their own files.

1. 1 Sequence in FASTA format

The FASTA format file is required for modules “AnnoGenome” and “AnaDiversity”.

FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The first line in a typical FASTA file started with a ">" as a header line. Following the initial line (used for a unique description of the sequence) is the actual sequence itself in standard one-letter character string. Anything other than a valid character would be ignored (including spaces, tabulators, asterisks, etc...).

1.2 Annotation reference in GenBank format

For the “AnnoGenome” module, if users plan to annotate their plastome using their own GenBank file as reference, a GenBank format file should be provided. The details for the description of GenBank file can be accessed at https://www.ncbi.nlm.nih.gov/genbank/.

3. Annotation in GFF format

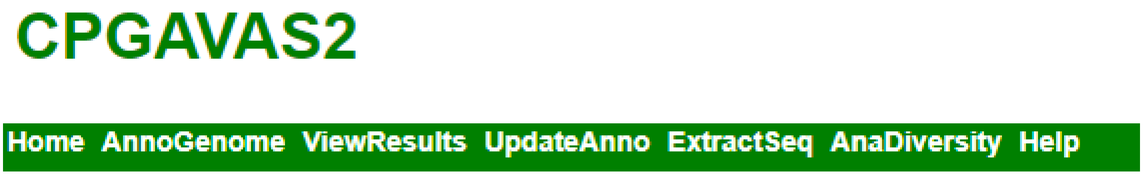
GFF format file is need for two modules “UpdateAnno” and “AnaDiversity”. For module “UpdateAnno”, the user can upload a manual curated GFF file using editors like Apollo. For module “AnaDiversity”, the GFF file is need for representing the reference annotations when running “RNA Editing Sites” pipeline. The details for the description of GFF file can be accessed at http://gmod.org/wiki/GFF.

4. FASTQ files

FASTQ files are needed when uses carry out “AnaDiversity” module, which including two pipelines: SNP discovery and RNA editing site discovery. At present, CPGVAS2 only accept NGS paired-end reads in FASTQ format generated by iIlumina sequencer. It should be pointed out, for the RNA-editing site analysis, the library type for sequencing should be “fr-firststrand”, which means the right-most end of the fragment (in transcript coordinates) is the first sequence. For multiple sets of pair-end sequencing results, users should merge all the left-end reads to one file and all the right-end reads into another files in the same order. The details for the explanation of FASAQ file can accessed at http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm.

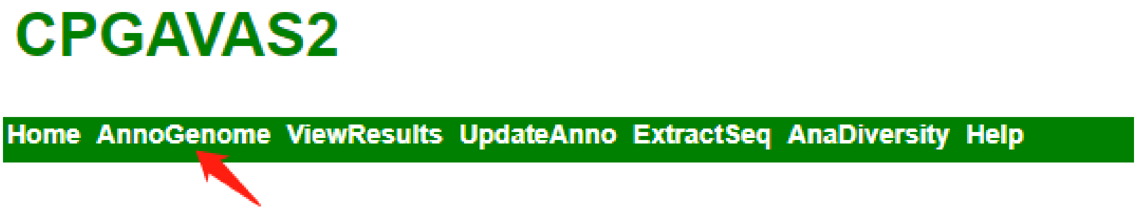
2 Guideline for using modules of CPGAVAS2

There are five modules provided by CPGAVAS2: a) AnnoGenome, b) ViewResults, c) UpdateAnno, d) ExtractSeq, e) AnaDiversity.

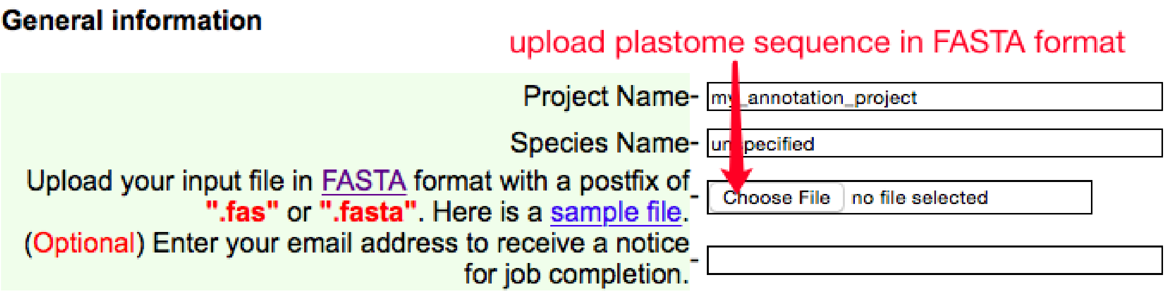


2.1 Module AnnoGenome: plastome sequence annotation

2.1.1 Go to module "AnnoGenome"



2.1.2 Upload your plastome sequence in FASTA format.



General information

upload plastome sequence in FASTA format

Project Name- my_annotation_project

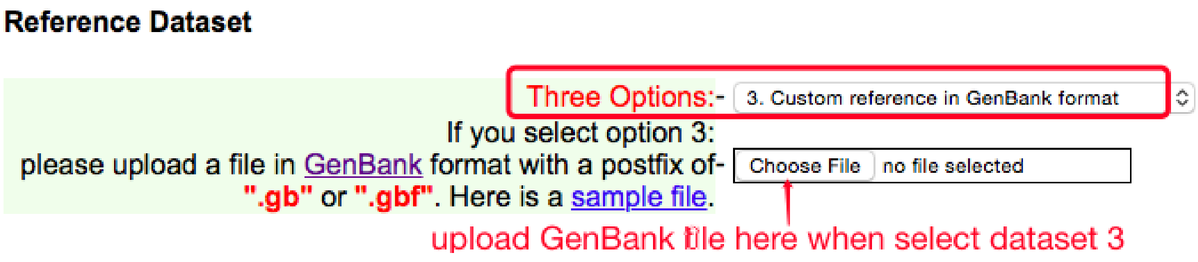
Species Name- unspecified

Upload your input file in FASTA format with a postfix of ".fas" or ".fasta". Here is a sample file. - Choose File no file selected

(Optional) Enter your email address to receive a notice for job completion. -

2.1.3 Select reference Dataset

For Reference Dataset, we provide three options. Most of time, the user would like to use 43-plastomes, which curated using RNA-seq data. These include all plastomes used in the datasets from GeSeq and DOGMA at genus level. However, 43 plastomes only represent a small fraction of plastome sequences that are currently available and might not contain the most closely related sequence for a particular query. At this time, the user can transfer to 2544. In the case when the user would like to use particular reference sequence, CPGAVAS2 allows the user to provide the sequence in GenBank format as reference.



Reference Dataset

Three Options:- 3. Custom reference in GenBank format

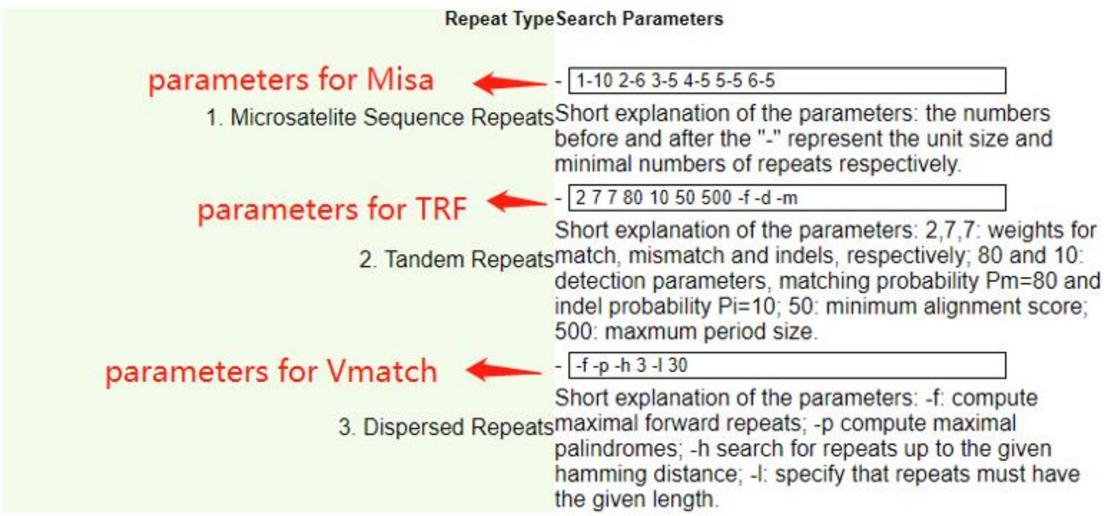
If you select option 3:
please upload a file in GenBank format with a postfix of- ".gb" or ".gbf". Here is a sample file.

Choose File no file selected

upload GenBank file here when select dataset 3

It should be point out, when you select “Custom reference in GenBank format” for the annotation, a reference file in GenBank format should be uploaded simultaneously.

2.1.4 The pipeline “Repeat identification” run simultaneously with plastome annotation. If you are very familiar with parameters for MISA, TRF and Vmatch, there are a lot of parameters can be adjusted. However, for routine analysis, we do not recommend user modify these parameters.



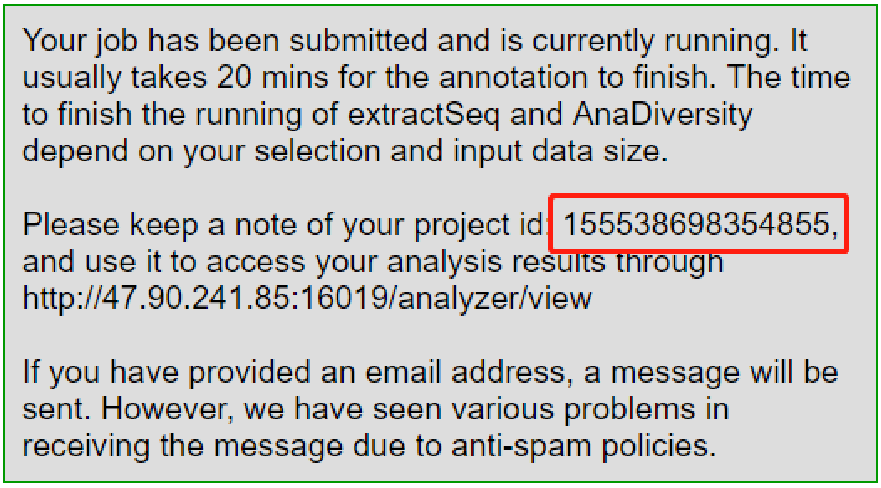
Repeat Type	Search Parameters
1. Microsatelite Sequence Repeats	- 1-10 2-6 3-5 4-5 5-5 6-5 Short explanation of the parameters: the numbers before and after the "-" represent the unit size and minimal numbers of repeats respectively.
2. Tandem Repeats	- 2 7 7 80 10 50 500 -f -d -m Short explanation of the parameters: 2,7,7: weights for match, mismatch and indels, respectively; 80 and 10: detection parameters, matching probability Pm=80 and indel probability Pi=10; 50: minimum alignment score; 500: maxmum period size.
3. Dispersed Repeats	- -f -p -h 3 -l 30 Short explanation of the parameters: -f: compute maximal forward repeats; -p compute maximal palindromes; -h search for repeats up to the given hamming distance; -l: specify that repeats must have the given length.

parameters for Misa

parameters for TRF

parameters for Vmatch

2.1.5 Click the "Submit" button and you will see a message (Figure 5). The best practice is to keep a note of the job ID and comes back 20 minutes later to check the results in "ViewResults" page.



Your job has been submitted and is currently running. It usually takes 20 mins for the annotation to finish. The time to finish the running of extractSeq and AnaDiversity depend on your selection and input data size.

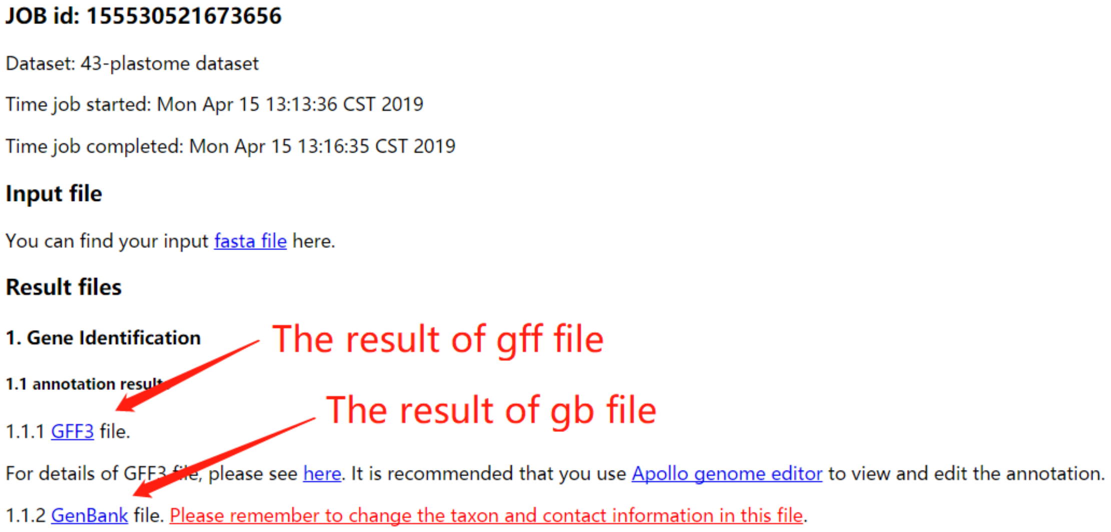
Please keep a note of your project id 155538698354855, and use it to access your analysis results through http://47.90.241.85:16019/analyzer/view

If you have provided an email address, a message will be sent. However, we have seen various problems in receiving the message due to anti-spam policies.

2.1.6 View and downloaded annotation results

Go to the "ViewResults" page, enter the project ID of your job to check the annotation results. A report page will show up. Browse through this page to see the results for 1) Gene Identification, include annotation results, sequences of the predicted genes and proteins, and analysis reports, 2) repeat elements.

1) Gene Identification in GFF file and GenBank file.



JOB id: 155530521673656

Dataset: 43-plastome dataset

Time job started: Mon Apr 15 13:13:36 CST 2019

Time job completed: Mon Apr 15 13:16:35 CST 2019

Input file

You can find your input fasta file here.

Result files

1. Gene Identification

The result of gff file

1.1 annotation result

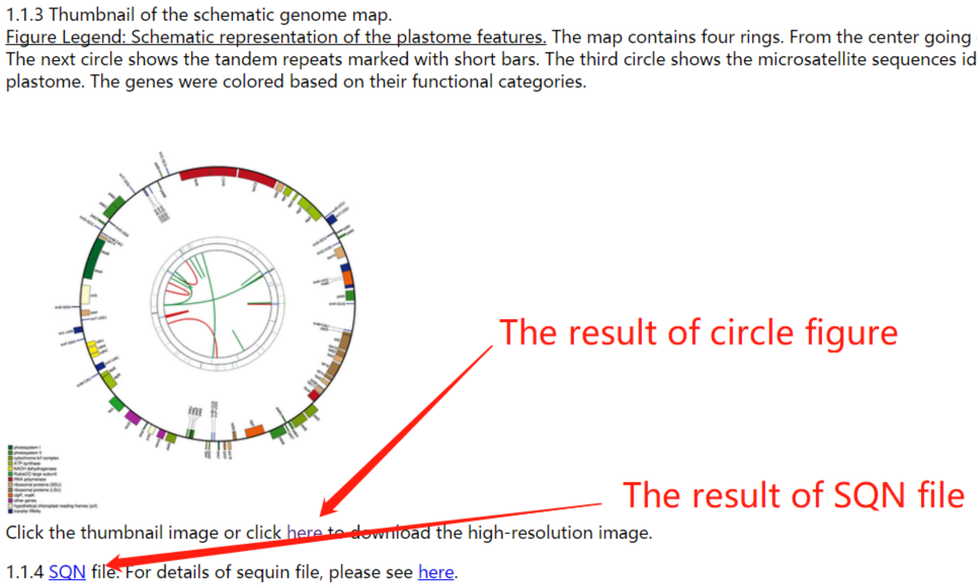
The result of gb file

1.1.1 GFF3 file.

For details of GFF3 file, please see here. It is recommended that you use Apollo genome editor to view and edit the annotation.

1.1.2 GenBank file. Please remember to change the taxon and contact information in this file.

2) Circular map and SQN file



1.1.3 Thumbnail of the schematic genome map.

Figure Legend: Schematic representation of the plastome features. The map contains four rings. From the center going outward, the first The next circle shows the tandem repeats marked with short bars. The third circle shows the microsatellite sequences identified using MI plastome. The genes were colored based on their functional categories.

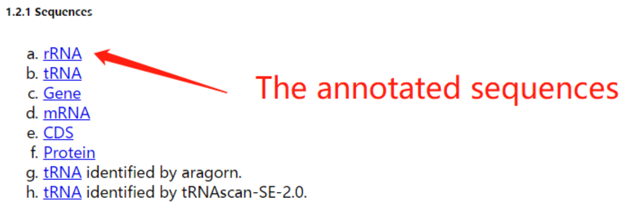
The result of circle figure

The result of SQN file

Click the thumbnail image or click here to download the high-resolution image.

1.1.4 SQN file. For details of sequin file, please see here.

3) Sequences. At this positon, CPGAVAS2 predicted rRNA, tRNA, Gene, mRNA, CDS, Proteins extracted from GenBank file produced by AnnoGenome module, if you click the corresponding blue word, the result will be opened or automatically downloaded depending your browser seting.

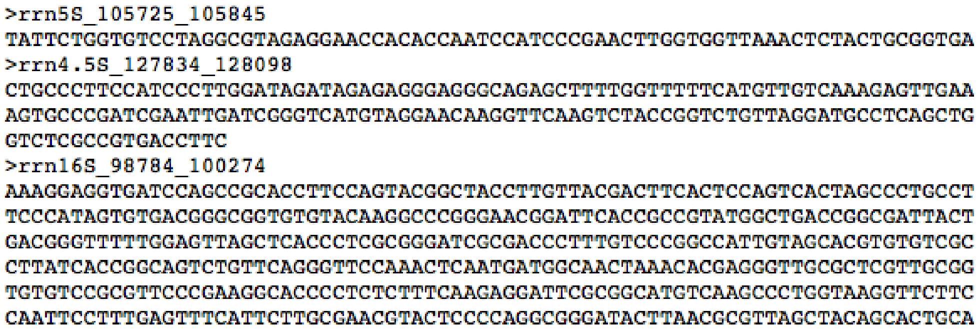


1.2.1 Sequences

a. rRNA
b. tRNA
c. Gene
d. mRNA
e. CDS
f. Protein
g. tRNA identified by aragorn.
h. tRNA identified by tRNAscan-SE-2.0.

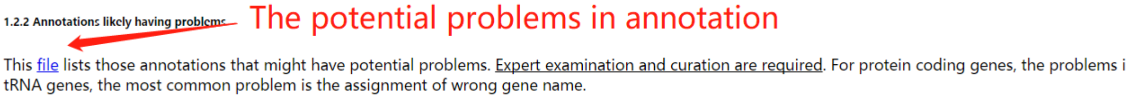
The annotated sequences

Here is an example file for rRNA sequences



```
>rrn5S_105725_105845
TATTCTGGTGTCCTAGGCGTAGAGGAACCACACCAATCCATCCCGAACTTGGTGGTTAAACTCTACTGCGGTGA
>rrn4.5S_127834_128098
CTGCCCTTCCATCCCTTGGATAGATAGAGAGGGAGGGCAGAGCTTTTGGTTTTTCATGTTGTCAAAGAGTTGAA
AGTGCCCGATCGAATTGATCGGGTCATGTAGGAACAAGGTTCAAGTCTACCGGTCTGTTAGGATGCCTCAGCTG
GTCTCGCCGTGACCTTC
>rrn16S_98784_100274
AAAGGAGGTGATCCAGCCGCACCTTCCAGTACGGCTACCTTGTTACGACTTCACTCCAGTCACTAGCCCTGCCT
TCCCATAGTGTGACGGGCGGTGTGTACAAGGCCCGGGAACGGATTCACCGCCGTATGGCTGACCGGCGATTACT
GACGGGTTTTTGGAGTTAGCTCACCCTCGCGGGATCGCGACCCTTTGTCCCGGCCATTGTAGCACGTGTGTCGC
CTTATCACCGGCAGTCTGTTCAGGGTTCCAAACTCAATGATGGCAACTAAACACGAGGGTTGCGCTCGTTGCGG
TGTGTCCGCGTTCCCGAAGGCACCCCTCTCTTTCAAGAGGATTCGCGGCATGTCAAGCCCTGGTAAGGTTCTTC
CAATTCCTTTGAGTTTCATTCTTGCGAACGTACTCCCCAGGCGGGATACTTAACGCGTTAGCTACAGCACTGCA
```

4) Annotations likely having problems.



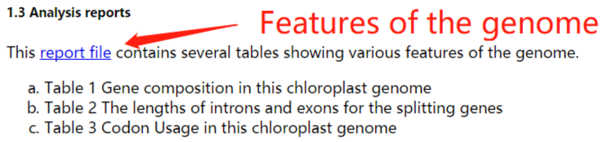
1.2.2 Annotations likely having problems

The potential problems in annotation

This file lists those annotations that might have potential problems. Expert examination and curation are required. For protein coding genes, the problems i tRNA genes, the most common problem is the assignment of wrong gene name.

5) Report files. This section contains three analysis report tables: 1) Gene composition in this

chloroplast genome. 2) The lengths of introns and exons for the splitting genes. 3) Codon Usage in this chloroplast genome.



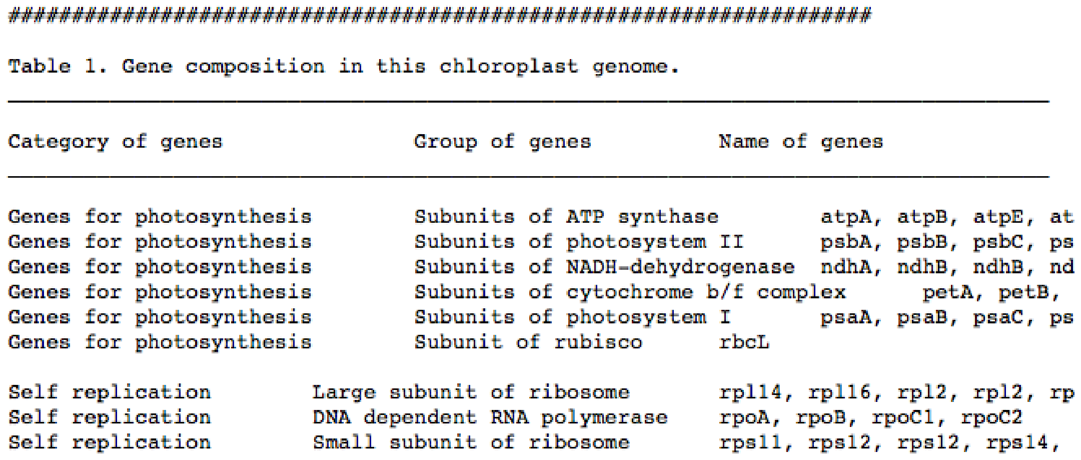
1.3 Analysis reports

Features of the genome

This report file contains several tables showing various features of the genome.

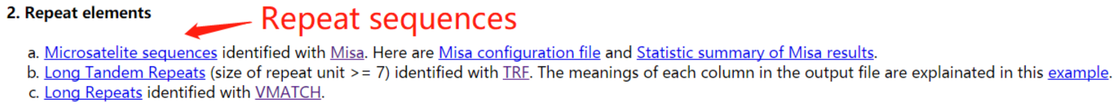
a. Table 1 Gene composition in this chloroplast genome
b. Table 2 The lengths of introns and exons for the splitting genes
c. Table 3 Codon Usage in this chloroplast genome

Here is an example for table 1



```
####################################################################

Table 1. Gene composition in this chloroplast genome.
___________________________________________________________________________________

Category of genes                Group of genes          Name of genes
___________________________________________________________________________________

Genes for photosynthesis         Subunits of ATP synthase       atpA, atpB, atpE, at
Genes for photosynthesis         Subunits of photosystem II     psbA, psbB, psbC, ps
Genes for photosynthesis         Subunits of NADH-dehydrogenase ndhA, ndhB, ndhB, nd
Genes for photosynthesis         Subunits of cytochrome b/f complex      petA, petB,
Genes for photosynthesis         Subunits of photosystem I      psaA, psaB, psaC, ps
Genes for photosynthesis         Subunit of rubisco      rbcL

Self replication        Large subunit of ribosome       rpl14, rpl16, rpl2, rpl2, rp
Self replication        DNA dependent RNA polymerase    rpoA, rpoB, rpoC1, rpoC2
Self replication        Small subunit of ribosome       rps11, rps12, rps12, rps14,
```

6) Repeat elements.

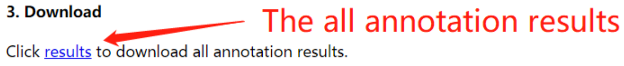


2. Repeat elements

Repeat sequences

a. Microsatelite sequences identified with Misa. Here are Misa configuration file and Statistic summary of Misa results.
b. Long Tandem Repeats (size of repeat unit >= 7) identified with TRF. The meanings of each column in the output file are explainated in this example.
c. Long Repeats identified with VMATCH.

7) downloaded results in “tar.gz” file



3. Download

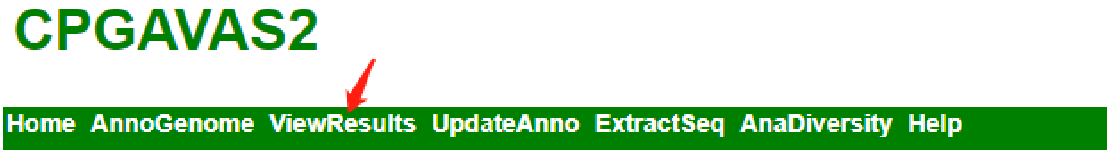
The all annotation results

Click results to download all annotation results.

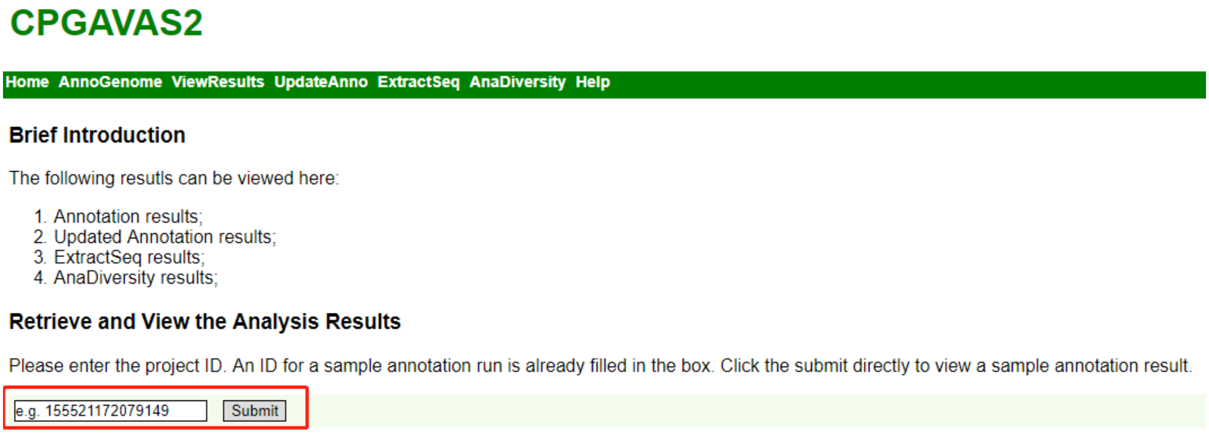
2.2 Module ViewResults: view results generated by CPGAVAS2

This module allows a user to view results generated by CPGAVAS2, including: Annotation results, Updated Annotation results, ExtractSeq results, and AnaDiversity results.

2.2.1 Go to module “ViewResults”


CPGAVAS2
Home AnnoGenome ViewResults UpdateAnno ExtractSeq AnaDiversity Help

2.2.2 Enter the project ID and click the submit button directly to view a sample annotation result.


CPGAVAS2
Home AnnoGenome ViewResults UpdateAnno ExtractSeq AnaDiversity Help

Brief Introduction

The following resutls can be viewed here:

1. Annotation results;
2. Updated Annotation results;
3. ExtractSeq results;
4. AnaDiversity results;

Retrieve and View the Analysis Results

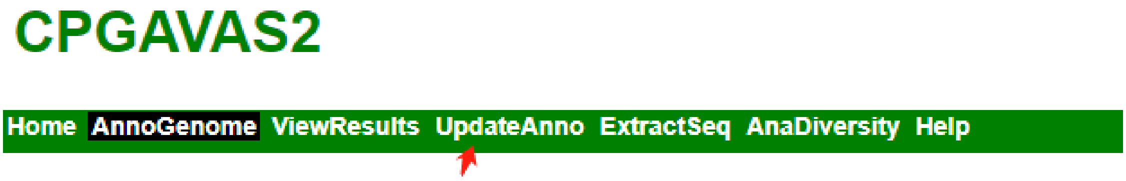
Please enter the project ID. An ID for a sample annotation run is already filled in the box. Click the submit directly to view a sample annotation result.

e.g. 155521172079149
Submit

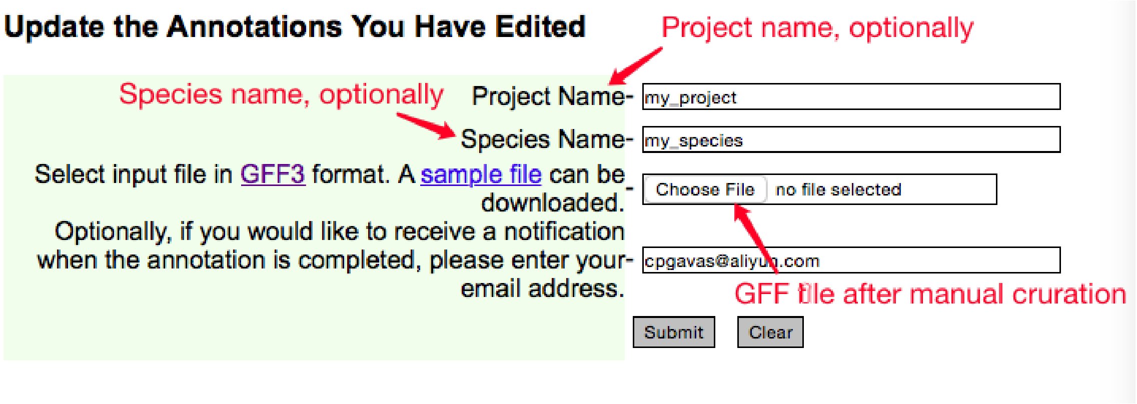
2.3 Module UpdateAnno: update annotation after manual curation

This module allows a user to update the annotation result after manual curation and recreate circular map and other analysis result.

2.3.1 Go to module “UpdateAnno”



2.3.2 Update the Annotations You have edited



Update the Annotations You Have Edited

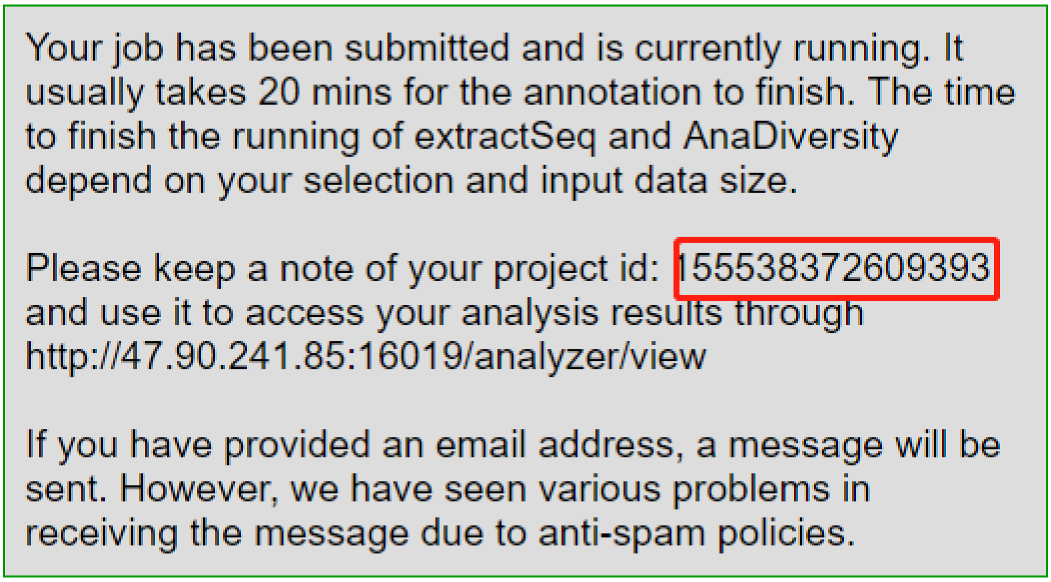
Project Name- my_project

Species Name- my_species

Select input file in GFF3 format. A sample file can be downloaded.-

Optionally, if you would like to receive a notification when the annotation is completed, please enter your email address.- cpgavas@aliyun.com

2.3.3 After the user clicks the “Submit” button, a message like the one below will pop up.



Your job has been submitted and is currently running. It usually takes 20 mins for the annotation to finish. The time to finish the running of extractSeq and AnaDiversity depend on your selection and input data size.

Please keep a note of your project id: 155538372609393 and use it to access your analysis results through http://47.90.241.85:16019/analyzer/view

If you have provided an email address, a message will be sent. However, we have seen various problems in receiving the message due to anti-spam policies.

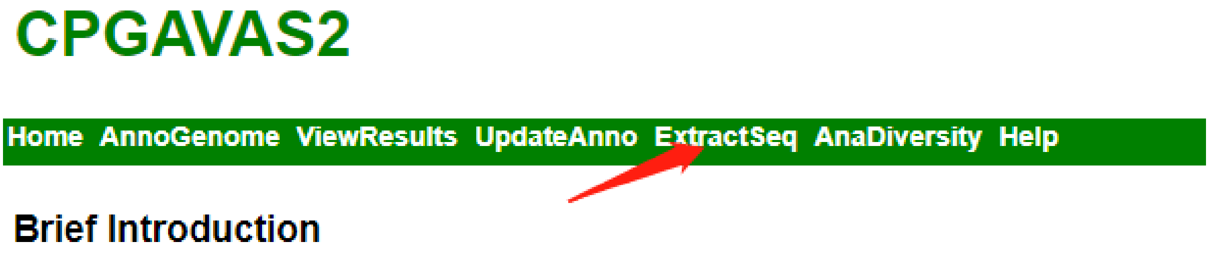
2.3.4 Results can be retrieved in the “ViewResults” module, output results are much similar that with “AnnoGenome”.

2.4 Module ExtractSeq: retrieve sequences

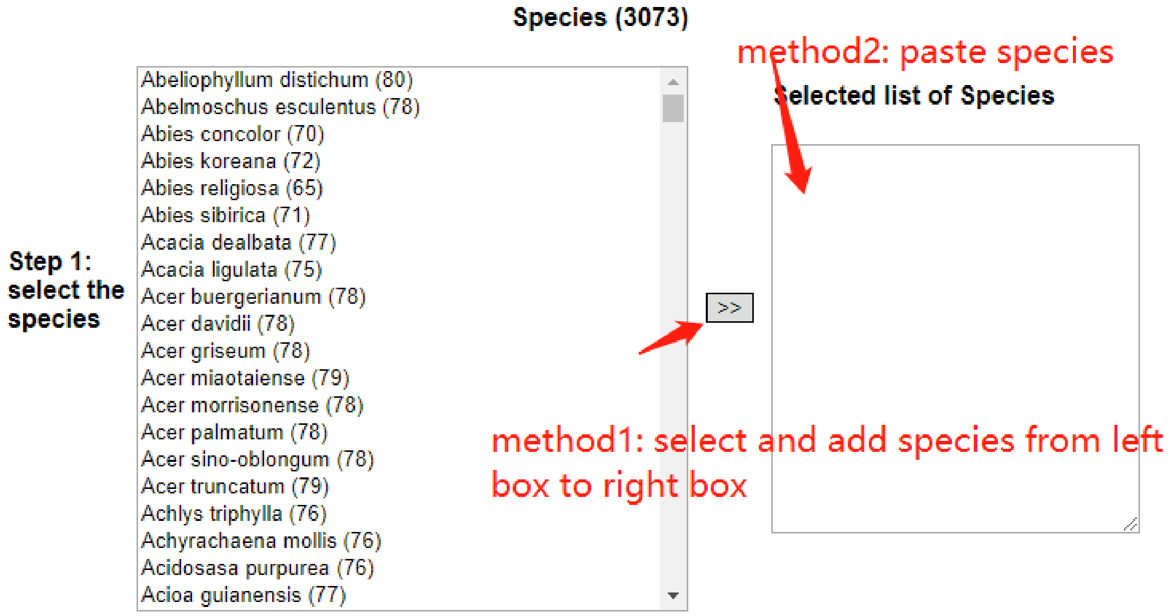
The module “ExtractSeq” is designed for retrieving CDS and protein sequences of public plastome sequences from NCBI “Refseq release of plastid sequences” database. This module can help users complete their phylogenetic analyses to understand taxonomic relationship between the newly obtained plastome and those having aleady been sequences.

This page allows users to retrieve protein and CDS sequences for lists of given genes and species name. Sequences will be provided in two formats, concatenated or non-concatenated, which can be subjected to phylogenetic analyses using either super-gene or super-tree methods. There two way to create species and gene lists that are used to retrieve sequences using ExtractSeq. The first method selects species or genes from the left boxes and then adds them to the box on the right. The second method allows users pastes species and gene names in the box on the right directly. Using the second method would cause the extraction to fail, especially user provide gene or species name inconsistent with the left boxes. Please pay special attention that the maximal numbers of species and genes are 10 and 80 respectively at one time.

1. Go to "ExtractSeq" page



2. Create species list



Species (3073)

Step 1: select the species

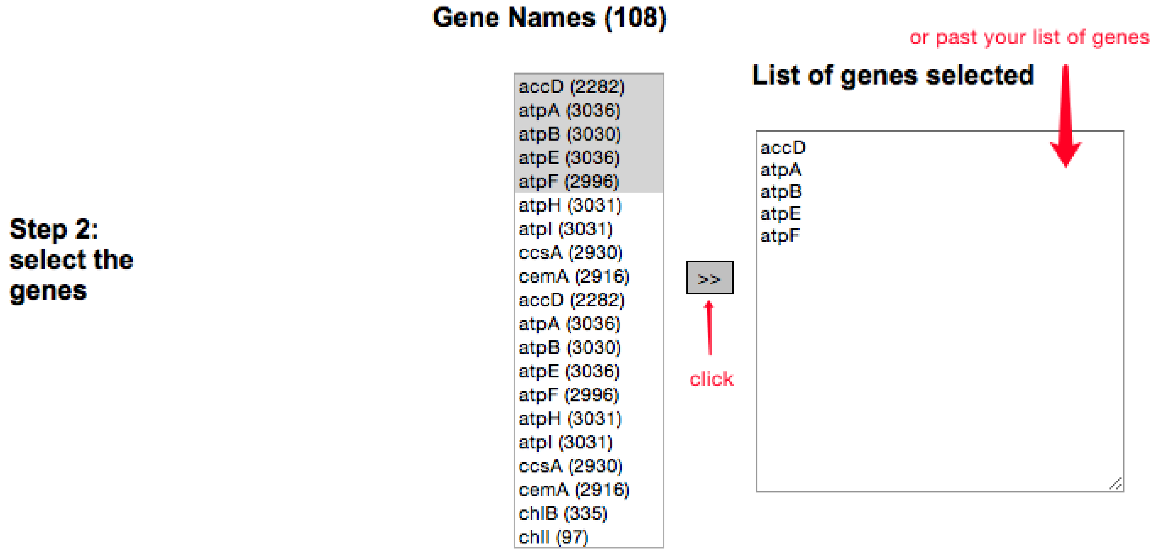
Abeliophyllum distichum (80)
Abelmoschus esculentus (78)
Abies concolor (70)
Abies koreana (72)
Abies religiosa (65)
Abies sibirica (71)
Acacia dealbata (77)
Acacia ligulata (75)
Acer buergerianum (78)
Acer davidii (78)
Acer griseum (78)
Acer miaotaiense (79)
Acer morrisonense (78)
Acer palmatum (78)
Acer sino-oblongum (78)
Acer truncatum (79)
Achlys triphylla (76)
Achyrachaena mollis (76)
Acidosasa purpurea (76)
Acioa guianensis (77)

Selected list of Species

method2: paste species

method1: select and add species from left box to right box

3. Create gene list



Gene Names (108)

Step 2: select the genes

accD (2282)
atpA (3036)
atpB (3030)
atpE (3036)
atpF (2996)
atpH (3031)
atpI (3031)
ccsA (2930)
cemA (2916)
accD (2282)
atpA (3036)
atpB (3030)
atpE (3036)
atpF (2996)
atpH (3031)
atpI (3031)
ccsA (2930)
cemA (2916)
chlB (335)
chlI (97)

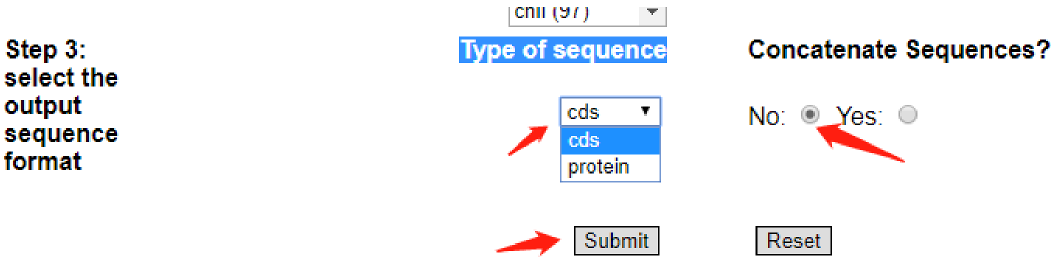
List of genes selected

accD
atpA
atpB
atpE
atpF

or past your list of genes

click

4. Select “Type of sequence” and their display method



Step 3: select the output sequence format

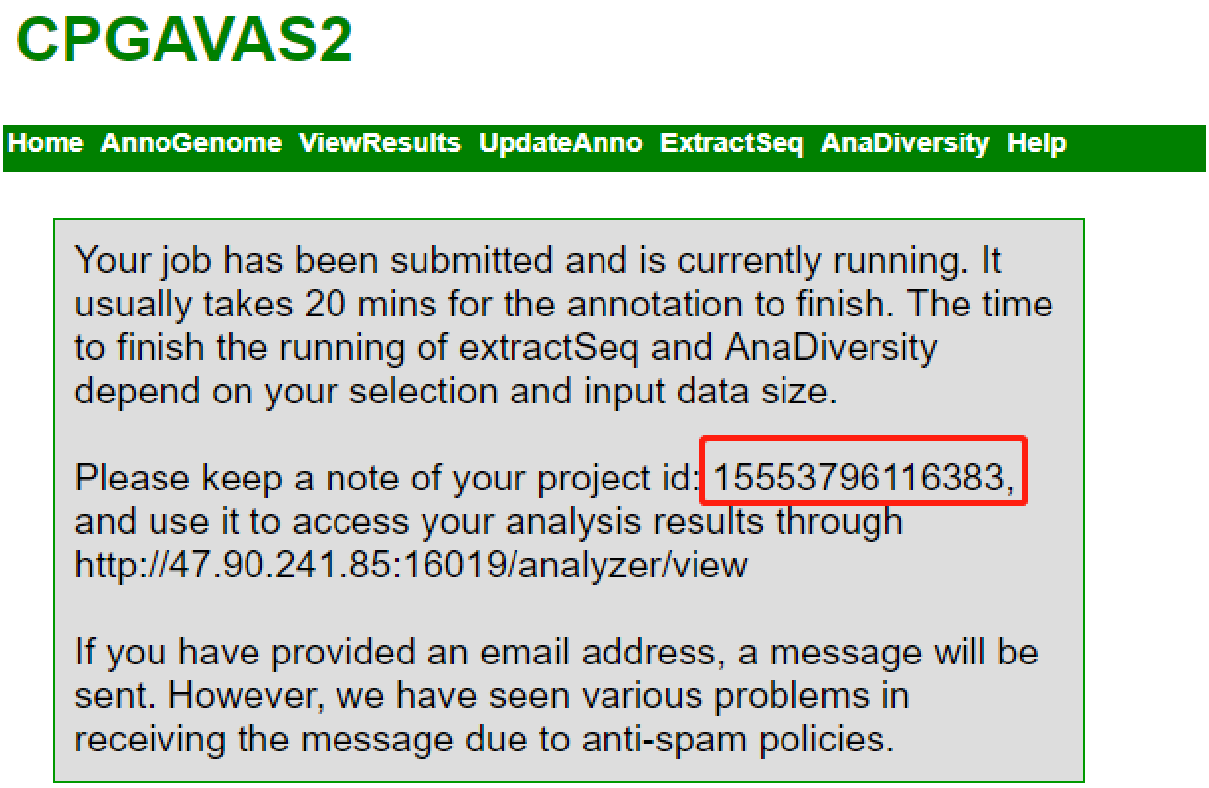
Type of sequence

cds
cds
protein

Concatenate Sequences?

No: ● Yes: ○

5. After clicking the "Submit" button and you will see a message like below. We encourage user copy and paste the Job id in a note pad.

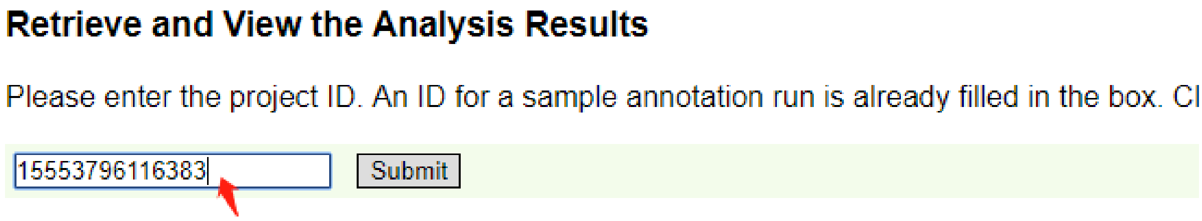


Your job has been submitted and is currently running. It usually takes 20 mins for the annotation to finish. The time to finish the running of extractSeq and AnaDiversity depend on your selection and input data size.

Please keep a note of your project id: 15553796116383, and use it to access your analysis results through http://47.90.241.85:16019/analyzer/view

If you have provided an email address, a message will be sent. However, we have seen various problems in receiving the message due to anti-spam policies.

6. Results can be viewed using the project id

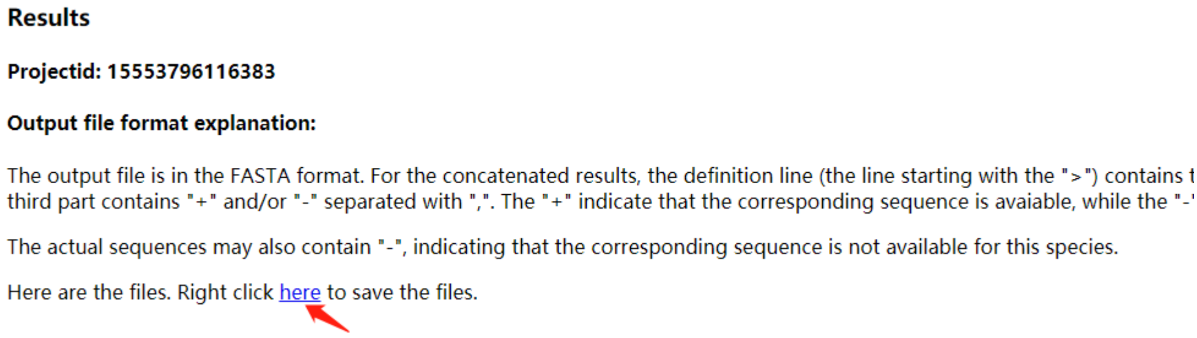


Retrieve and View the Analysis Results

Please enter the project ID. An ID for a sample annotation run is already filled in the box. Cl

15553796116383

7. Right click and save your result



Results

Projectid: 15553796116383

Output file format explanation:

The output file is in the FASTA format. For the concatenated results, the definition line (the line starting with the ">") contains t third part contains "+" and/or "-" separated with ",". The "+" indicate that the corresponding sequence is avaiable, while the "-'

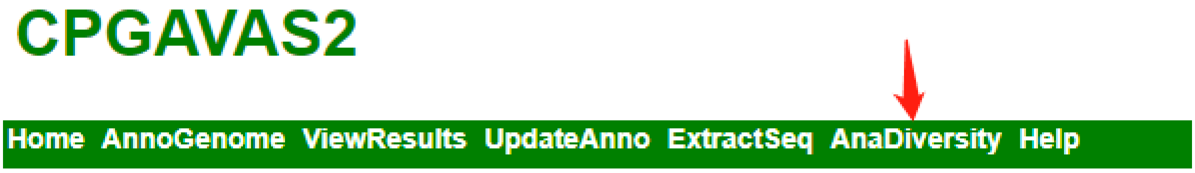
The actual sequences may also contain "-", indicating that the corresponding sequence is not available for this species.

Here are the files. Right click here to save the files.

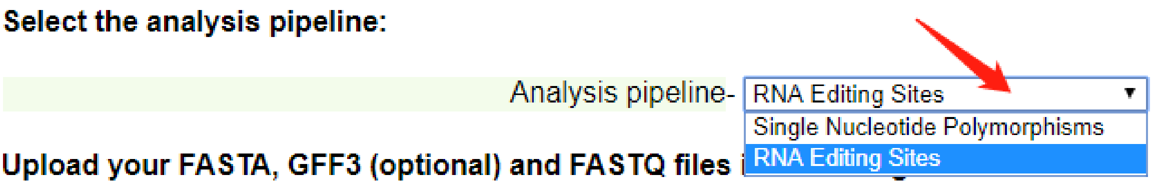
2.5 Module AnaDiversity: discovery SNP and RNA-editing Site

This page allows users to discovery SNP and RNA-editing Site from species.

1. Go to "AnaDiversity" page.



2. Select the analysis pipeline “Single Nucleotide Polymorphisms or RNA Editing Sites”

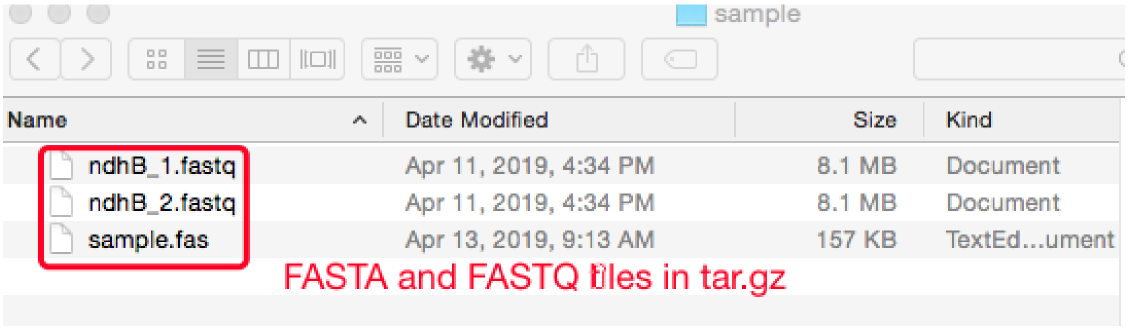


Select the analysis pipeline:

Analysis pipeline- RNA Editing Sites

Upload your FASTA, GFF3 (optional) and FASTQ files i

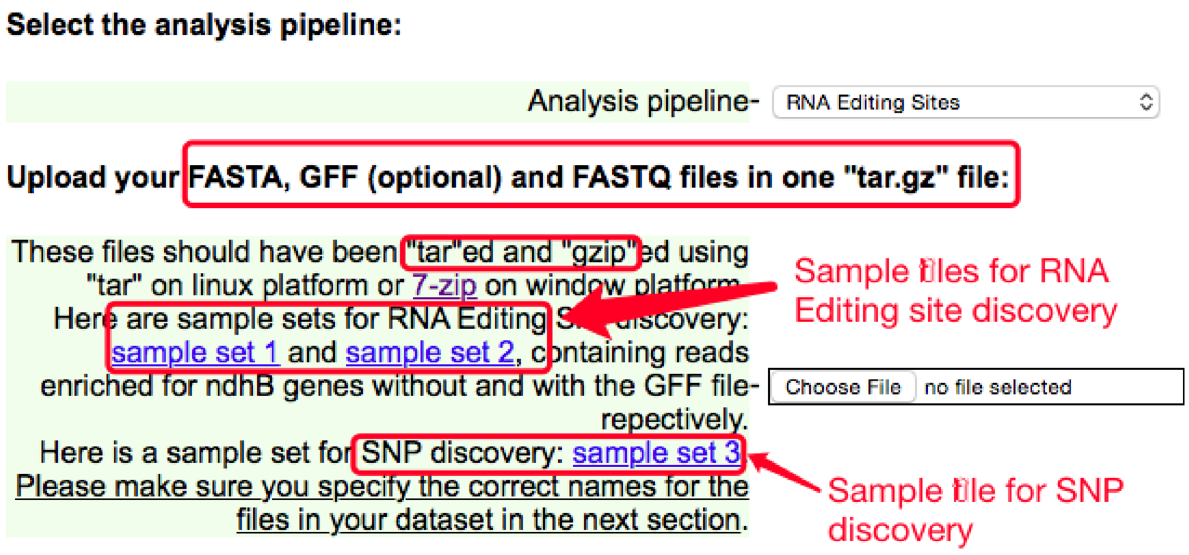
3. Prepare your dataset for diversity analysis in one "tar.gz" file. For “Single Nucleotide Polymorphisms” pipeline, the reference sequence in FASTA format and NGS sequences reads in FASTA format are needed. For “RNA Editing Sites” pipeline, in addition to the above two files, you can also choose to submit a GFF file for representing the reference annotations.



Name	Date Modified	Size	Kind
ndhB_1.fastq	Apr 11, 2019, 4:34 PM	8.1 MB	Document
ndhB_2.fastq	Apr 11, 2019, 4:34 PM	8.1 MB	Document
sample.fas	Apr 13, 2019, 9:13 AM	157 KB	TextEd...ument

FASTA and FASTQ files in tar.gz

User can use our sample file for testing, download the Sample set 1 and Sample set 2 are for RNA-editing and sample set 3 was for SNP.



Select the analysis pipeline:

Analysis pipeline- RNA Editing Sites

Upload your FASTA, GFF (optional) and FASTQ files in one "tar.gz" file:

These files should have been "tar"ed and "gzip"ed using "tar" on linux platform or 7-zip on window platform. Here are sample sets for RNA Editing Site discovery: sample set 1 and sample set 2, containing reads enriched for ndhB genes without and with the GFF file- repectively.

Here is a sample set for SNP discovery: sample set 3

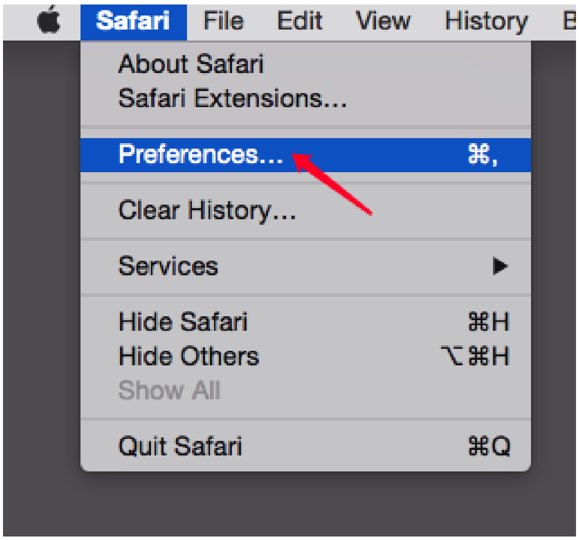
Please make sure you specify the correct names for the files in your dataset in the next section.

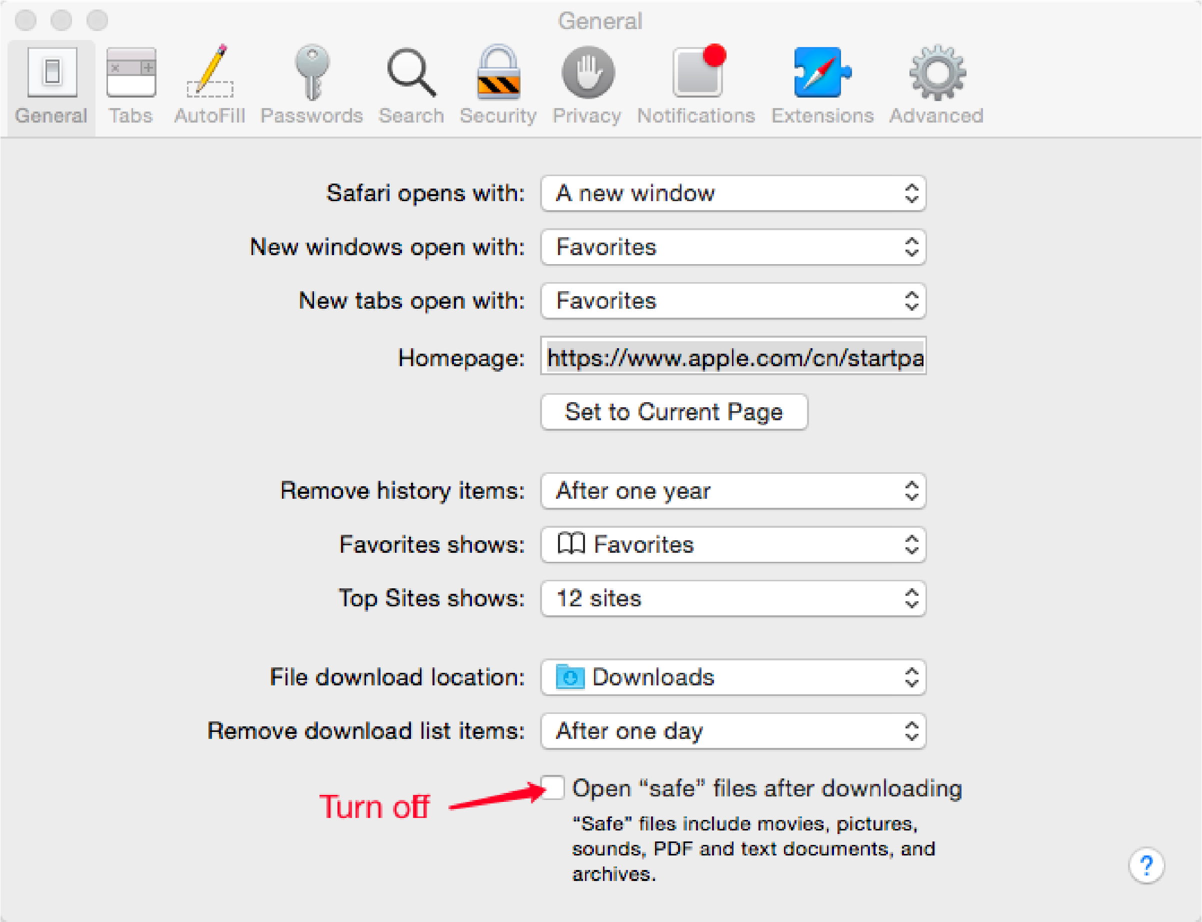
Choose File no file selected

Sample files for RNA Editing site discovery

Sample file for SNP discovery

If you use mac computer, the sample .tar.gz file would be changed to be .tar file automatically. At this time, you need to change your safari preferences following the instructions below. Briefly, from Safari->Preference, uncheck open “saft” file after downloading.





Safari opens with: A new window

New windows open with: Favorites

New tabs open with: Favorites

Homepage: https://www.apple.com/cn/startpa

Set to Current Page

Remove history items: After one year

Favorites shows: Favorites

Top Sites shows: 12 sites

File download location: Downloads

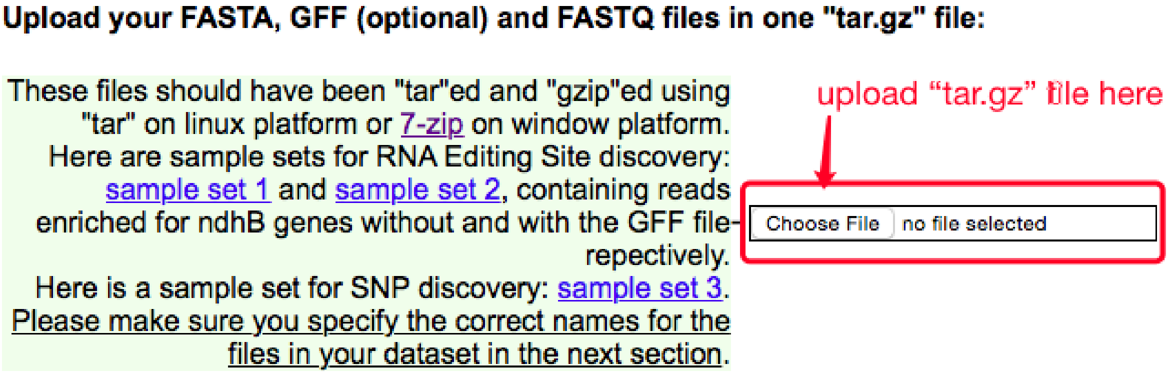
Remove download list items: After one day

Turn off

□ Open “safe” files after downloading

“Safe” files include movies, pictures, sounds, PDF and text documents, and archives.

4. Upload your FASTA, GFF3 (optional) and FASTQ files in one "tar.gz" file

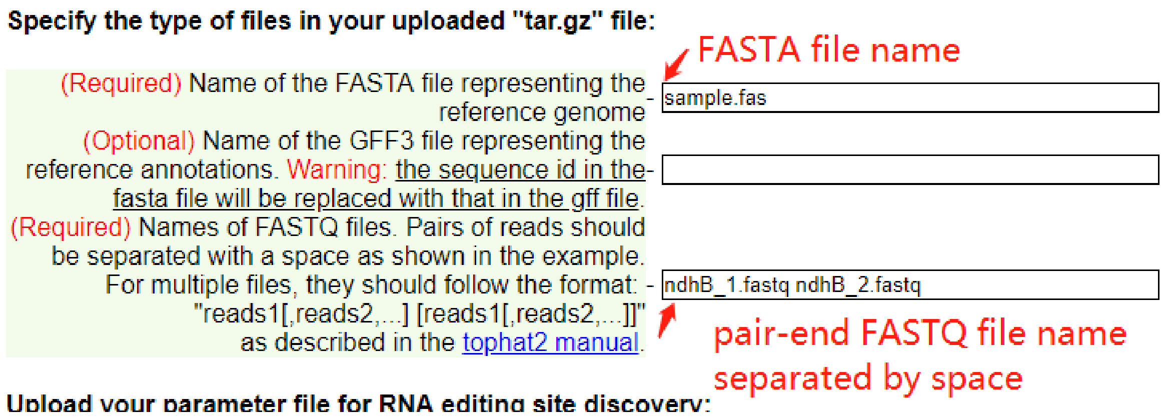


Upload your FASTA, GFF (optional) and FASTQ files in one "tar.gz" file:

These files should have been "tar"ed and "gzip"ed using "tar" on linux platform or 7-zip on window platform. Here are sample sets for RNA Editing Site discovery: sample set 1 and sample set 2, containing reads enriched for ndhB genes without and with the GFF file repectively. Here is a sample set for SNP discovery: sample set 3. Please make sure you specify the correct names for the files in your dataset in the next section.

upload “tar.gz” file here

5. Specify names for each type of files in your uploaded "tar.gz" file:



Specify the type of files in your uploaded "tar.gz" file:

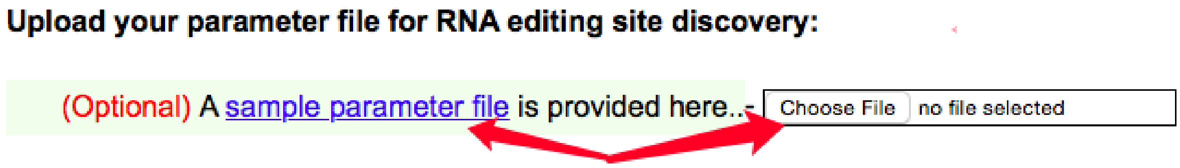
(Required) Name of the FASTA file representing the reference genome - sample.fas

(Optional) Name of the GFF3 file representing the reference annotations. Warning: the sequence id in the fasta file will be replaced with that in the gff file. -

(Required) Names of FASTQ files. Pairs of reads should be separated with a space as shown in the example. For multiple files, they should follow the format: "reads1[,reads2,...] [reads1[,reads2,...]]" as described in the tophat2 manual. - ndhB_1.fastq ndhB_2.fastq

Upload your parameter file for RNA editing site discovery:

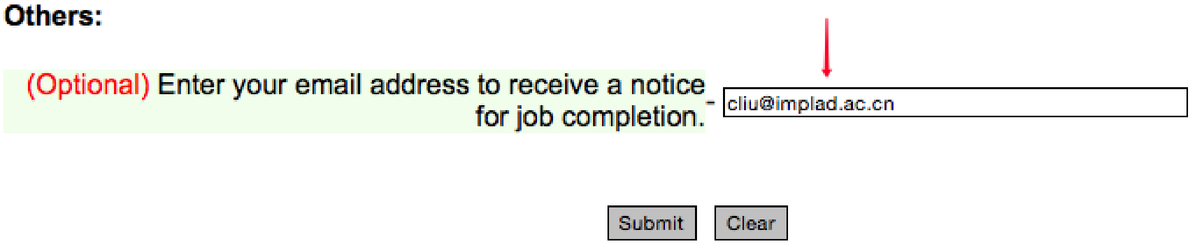
6. Optionally, there are a lot of parameters in a configure file for adjusting, however, upload a specific configure file is not encouraged especially for regular analysis.



Upload your parameter file for RNA editing site discovery:

(Optional) A sample parameter file is provided here.. - Choose File no file selected

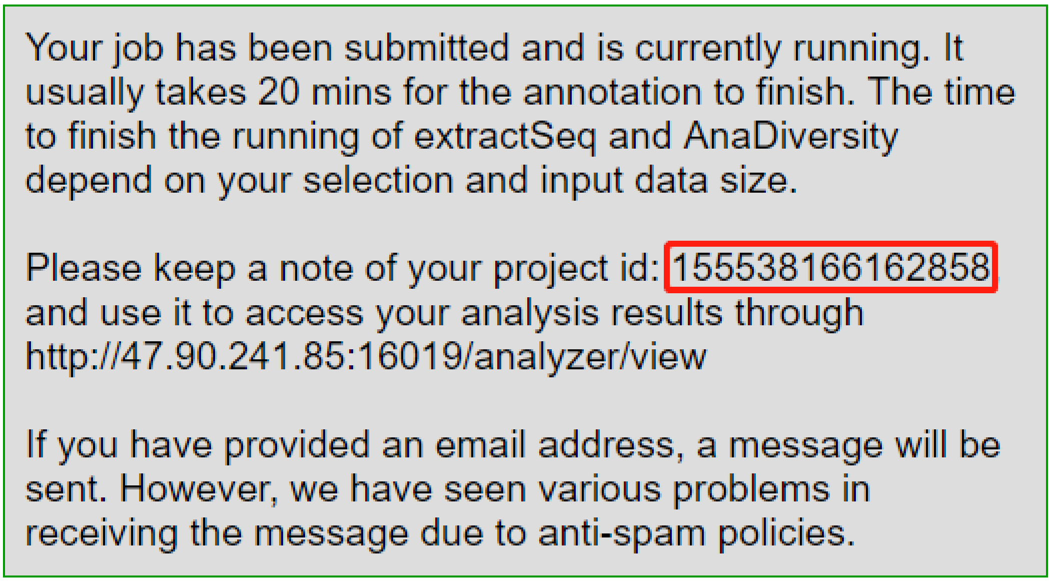
7. Optionally, enter your email address to receive a notice for job completion.



Others:

(Optional) Enter your email address to receive a notice for job completion. - cliu@implad.ac.cn

8. Click the "Submit" button and you will see a message like below. The best practice is to keep a note of the job ID and comes back 20 minutes later to check the results.

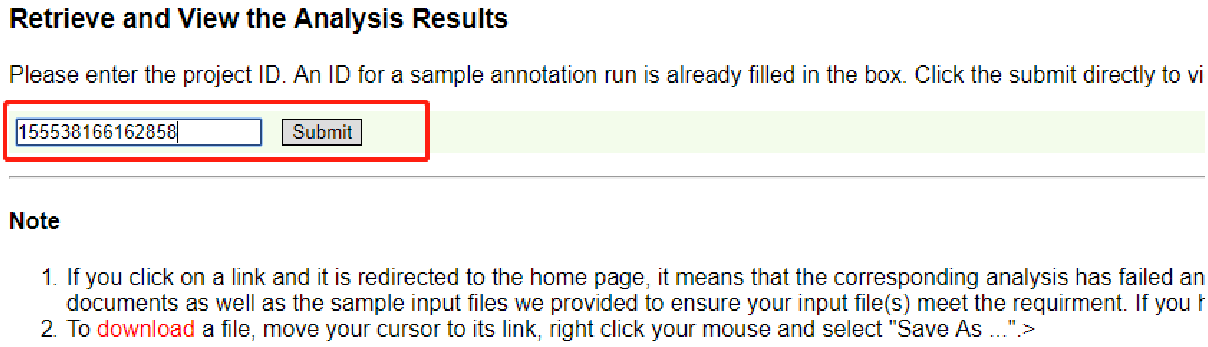


Your job has been submitted and is currently running. It usually takes 20 mins for the annotation to finish. The time to finish the running of extractSeq and AnaDiversity depend on your selection and input data size.

Please keep a note of your project id: 1555381661628858 and use it to access your analysis results through http://47.90.241.85:16019/analyzer/view

If you have provided an email address, a message will be sent. However, we have seen various problems in receiving the message due to anti-spam policies.

9. Once you come back, go to the "ViewResults" page, and enter the project ID of your job. A report page will show there.



Retrieve and View the Analysis Results

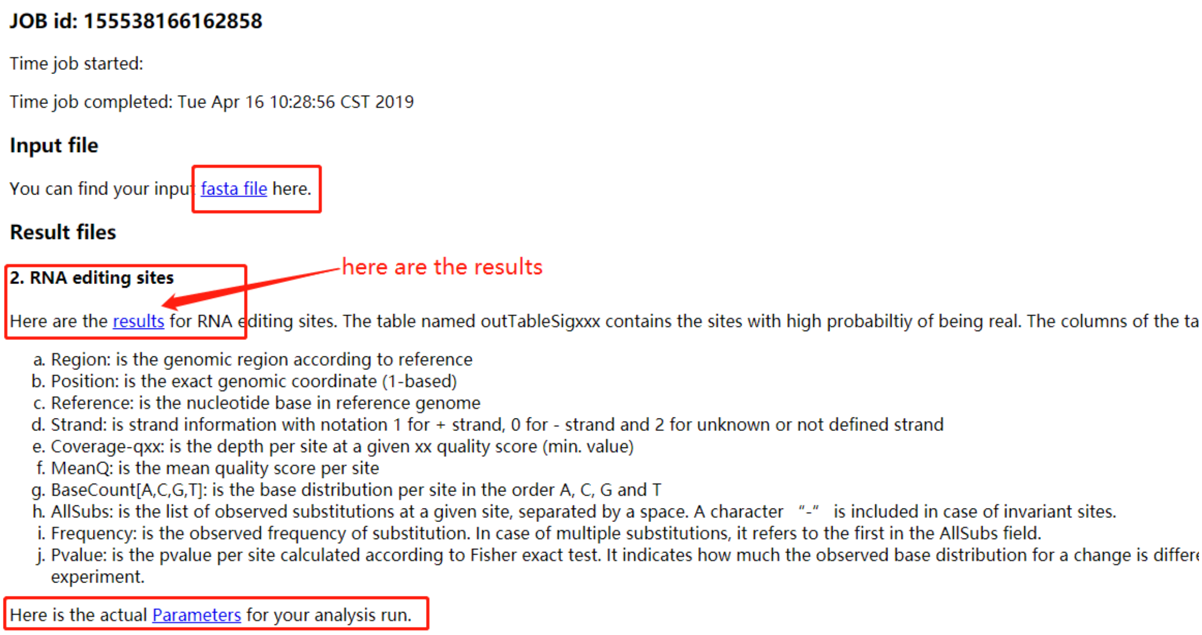
Please enter the project ID. An ID for a sample annotation run is already filled in the box. Click the submit directly to vi

155538166162858

Note

1. If you click on a link and it is redirected to the home page, it means that the corresponding analysis has failed an documents as well as the sample input files we provided to ensure your input file(s) meet the requirment. If you h
2. To download a file, move your cursor to its link, right click your mouse and select "Save As ...".>

10. Wight click and download the ".tar.gz" files to view your result.



JOB id: 155538166162858

Time job started:

Time job completed: Tue Apr 16 10:28:56 CST 2019

Input file

You can find your input fasta file here.

Result files

here are the results

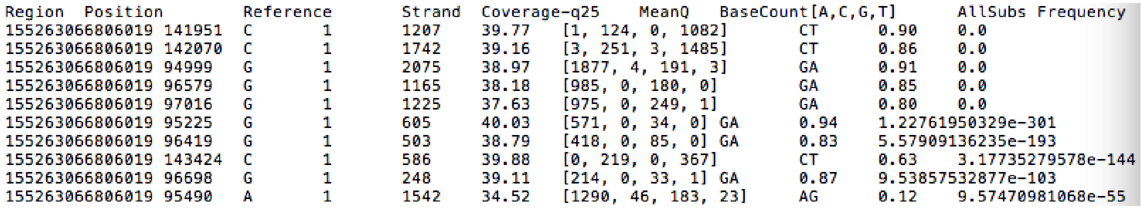
2. RNA editing sites

Here are the results for RNA editing sites. The table named outTableSigxxx contains the sites with high probabiltiy of being real. The columns of the ta

a. Region: is the genomic region according to reference
b. Position: is the exact genomic coordinate (1-based)
c. Reference: is the nucleotide base in reference genome
d. Strand: is strand information with notation 1 for + strand, 0 for - strand and 2 for unknown or not defined strand
e. Coverage-qxx: is the depth per site at a given xx quality score (min. value)
f. MeanQ: is the mean quality score per site
g. BaseCount[A,C,G,T]: is the base distribution per site in the order A, C, G and T
h. AllSubs: is the list of observed substitutions at a given site, separated by a space. A character "-" is included in case of invariant sites.
i. Frequency: is the observed frequency of substitution. In case of multiple substitutions, it refers to the first in the AllSubs field.
j. Pvalue: is the pvalue per site calculated according to Fisher exact test. It indicates how much the observed base distribution for a change is differe experiment.

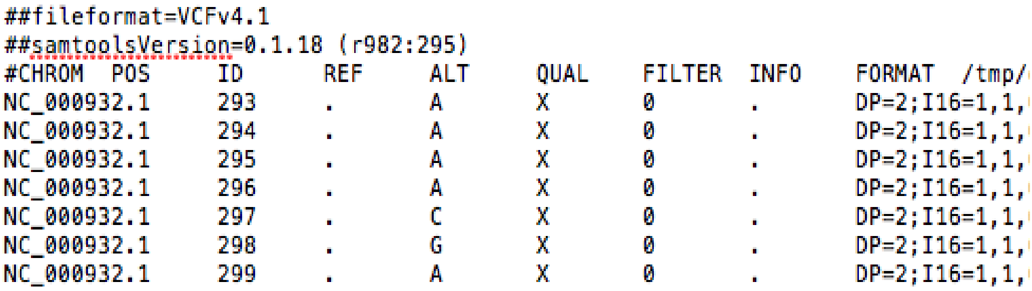
Here is the actual Parameters for your analysis run.

11. RNA editing site results has two .txt file, include forward and reverse RNA editing site massages



```
Region  Position        Reference       Strand  Coverage-q25    MeanQ   BaseCount[A,C,G,T]      AllSubs Frequency
155263066806019 141951  C       1       1207    39.77   [1, 124, 0, 1082]       CT      0.90    0.0
155263066806019 142070  C       1       1742    39.16   [3, 251, 3, 1485]       CT      0.86    0.0
155263066806019 94999   G       1       2075    38.97   [1877, 4, 191, 3]       GA      0.91    0.0
155263066806019 96579   G       1       1165    38.18   [985, 0, 180, 0]        GA      0.85    0.0
155263066806019 97016   G       1       1225    37.63   [975, 0, 249, 1]        GA      0.80    0.0
155263066806019 95225   G       1       605     40.03   [571, 0, 34, 0] GA      0.94    1.22761950329e-301
155263066806019 96419   G       1       503     38.79   [418, 0, 85, 0] GA      0.83    5.57909136235e-193
155263066806019 143424  C       1       586     39.88   [0, 219, 0, 367]        CT      0.63    3.17735279578e-144
155263066806019 96698   G       1       248     39.11   [214, 0, 33, 1] GA      0.87    9.53857532877e-103
155263066806019 95490   A       1       1542    34.52   [1290, 46, 183, 23]     AG      0.12    9.57470981068e-55
```

12. SNP sites results has one vcf file



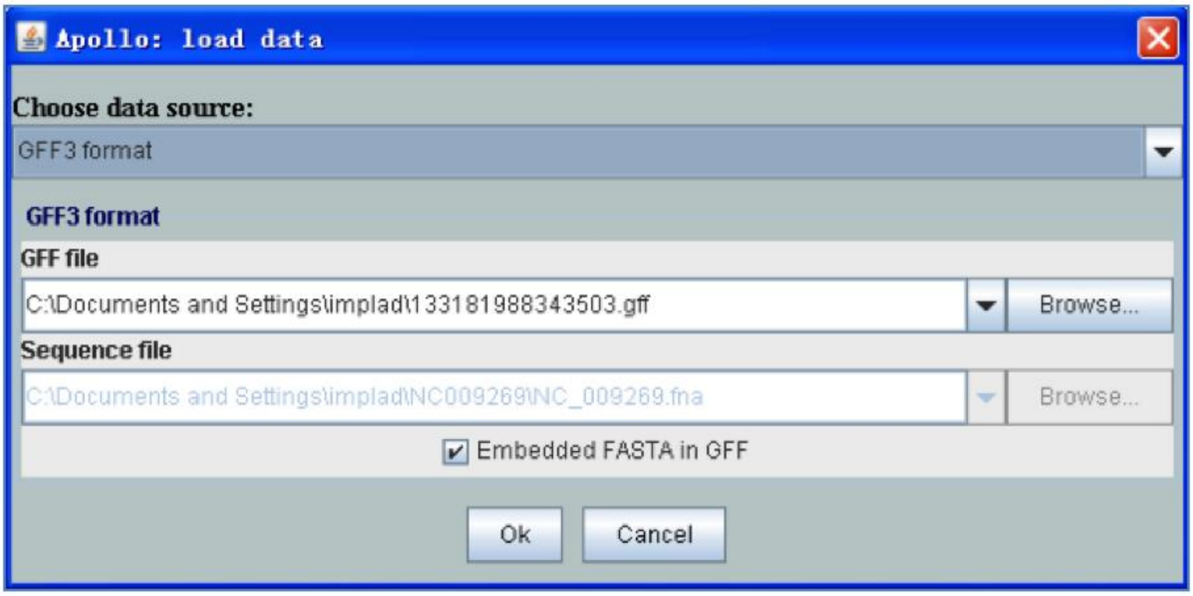
```
##fileformat=VCFv4.1
##samtoolsVersion=0.1.18 (r982:295)
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  /tmp/
NC_000932.1     293     .       A       X       0       .       DP=2;I16=1,1,
NC_000932.1     294     .       A       X       0       .       DP=2;I16=1,1,
NC_000932.1     295     .       A       X       0       .       DP=2;I16=1,1,
NC_000932.1     296     .       A       X       0       .       DP=2;I16=1,1,
NC_000932.1     297     .       C       X       0       .       DP=2;I16=1,1,
NC_000932.1     298     .       G       X       0       .       DP=2;I16=1,1,
NC_000932.1     299     .       A       X       0       .       DP=2;I16=1,1,
```

Appendix：A brief guideline for editing CPGAVAS2 annotation result in GFF file using Apollo

Apollo is a genome annotation-editing tool with an easy to use graphical interface. Here is a brief guideline to familiarize users to edit CPGAVAS2 annotation result in GFF file using Apollo, detailed documentation please connect to http://genomearchitect.org/users-guide/. We describe below some of the most commonly used tasks users need to perform to edit the annotations result in GFF file generated by CPGAVAS2. The original and complete tutorial can be found at http://genomearchitect.org/users-guide/.

I. Read GFF3 files

GFF3 format is used to represent the genomic annotations produced by **CPGAVAS**. To read in a GFF3 file, choose the "GFF3 format" option from the data adapter menu. Click the “Browse” button to bring up a file chooser. You must either provide the sequence data in FASTA format or you will need to have the FASTA data embedded in the GFF3 file. For the first option, make sure that "Embedded FASTA in GFF" is not checked and enter the FASTA file to go with the GFF3 data (you can browse for the file as well). For the second option, make sure that "Embedded FASTA in GFF" is checked. This will disable the "Sequence file" selection. When you're ready, press the OK button and the GFF3 file and sequence data will be read in and the features displayed.



Apollo: load data

Choose data source: GFF3 format

GFF3 format

GFF file: C:\Documents and Settings\implad\1331819883435O3.gff

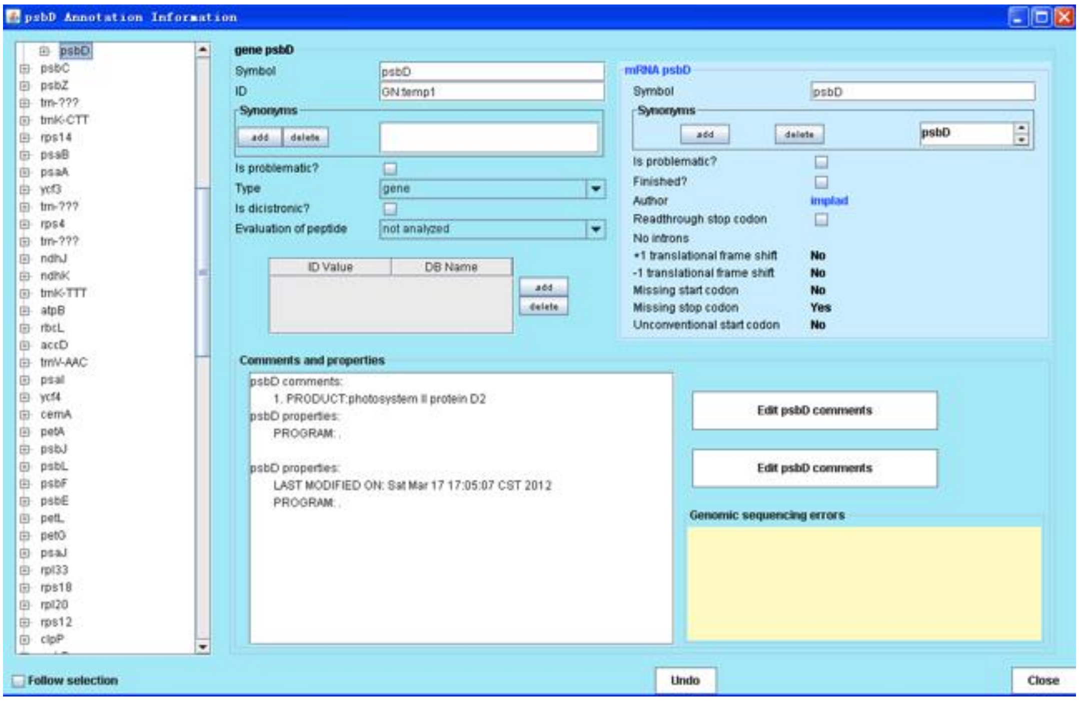
Sequence file: C:\Documents and Settings\implad\NC009269\NC_009269.fna

☑ Embedded FASTA in GFF

II. Invoke Edit Annotation Text Editor

To bring up the annotation text editor for a transcript, choose the "Annotation info editor" option in the right-mouse popup menu. The annotation information window will appear for the selected

gene/transcript, which allows the curator to 1) select a type of entity for the annotation, e.g. gene (protein-coding gene, the default), tRNA, transposable_element, etc.; 2) change the symbol, ID, or synonyms of the annotations or transcripts; or 3) add comments to annotations or individual transcripts.



psbD Annotation Information

gene psbD

Symbol: psbD

ID: GN:temp1

Is problematic? □

Type: gene

Is dicistronic? □

Evaluation of peptide: not analyzed

ID Value	DB Name

mRNA psbD

Symbol: psbD

Synonyms: psbD

Is problematic? □

Finished? □

Author: implad

Readthrough stop codon □

No introns

+1 translational frame shift	No
-1 translational frame shift	No
Missing start codon	No
Missing stop codon	Yes
Unconventional start codon	No

Comments and properties

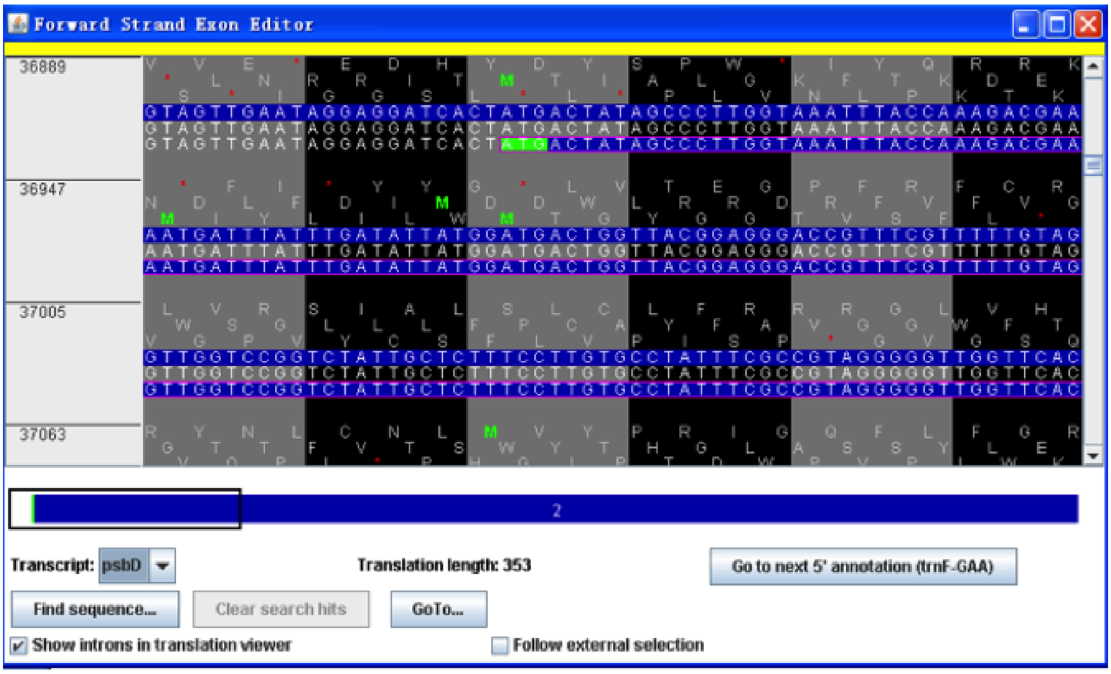
```
psbD comments:
    1. PRODUCT:photosystem II protein D2
psbD properties:
    PROGRAM: .

psbD properties:
    LAST MODIFIED ON: Sat Mar 17 17:05:07 CST 2012
    PROGRAM: .
```

Genomic sequencing errors

□ Follow selection

III. Invoke Exon Detail Editor

The Exon Detail Editor can be invoked from the right-mouse popup menu when you select an annotation that has sequence associated with it. A separate window will appear that shows the reference nucleotide sequence centered around the selected feature.


Forward Strand Exon Editor
Find sequence...
Clear search hits
GoTo...
Go to next 5' annotation (trnF-GAA)
2

```
36889  GTAGTTGAATAGGAGGATCACTATGACTATAGCCCTTGGTAAATTTACCAAAGACGAA
       GTAGTTGAATAGGAGGATCACTATGACTATAGCCCTTGGTAAATTTACCAAAGACGAA
       GTAGTTGAATAGGAGGATCACTATGACTATAGCCCTTGGTAAATTTACCAAAGACGAA
36947  AATGATTTATTTGATATTATGGATGACTGGTTACGGAGGGACCGTTTCGTTTTTGTAG
       AATGATTTATTTGATATTATGGATGACTGGTTACGGAGGGACCGTTTCGTTTTTGTAG
       AATGATTTATTTGATATTATGGATGACTGGTTACGGAGGGACCGTTTCGTTTTTGTAG
37005  GTTGGTCCGGTCTATTGCTCTTTCCTTGTGCCTATTTCGCCGTAGGGGGTTGGTTCAC
       GTTGGTCCGGTCTATTGCTCTTTCCTTGTGCCTATTTCGCCGTAGGGGGTTGGTTCAC
       GTTGGTCCGGTCTATTGCTCTTTCCTTGTGCCTATTTCGCCGTAGGGGGTTGGTTCAC
37063
```

Transcript: psbD

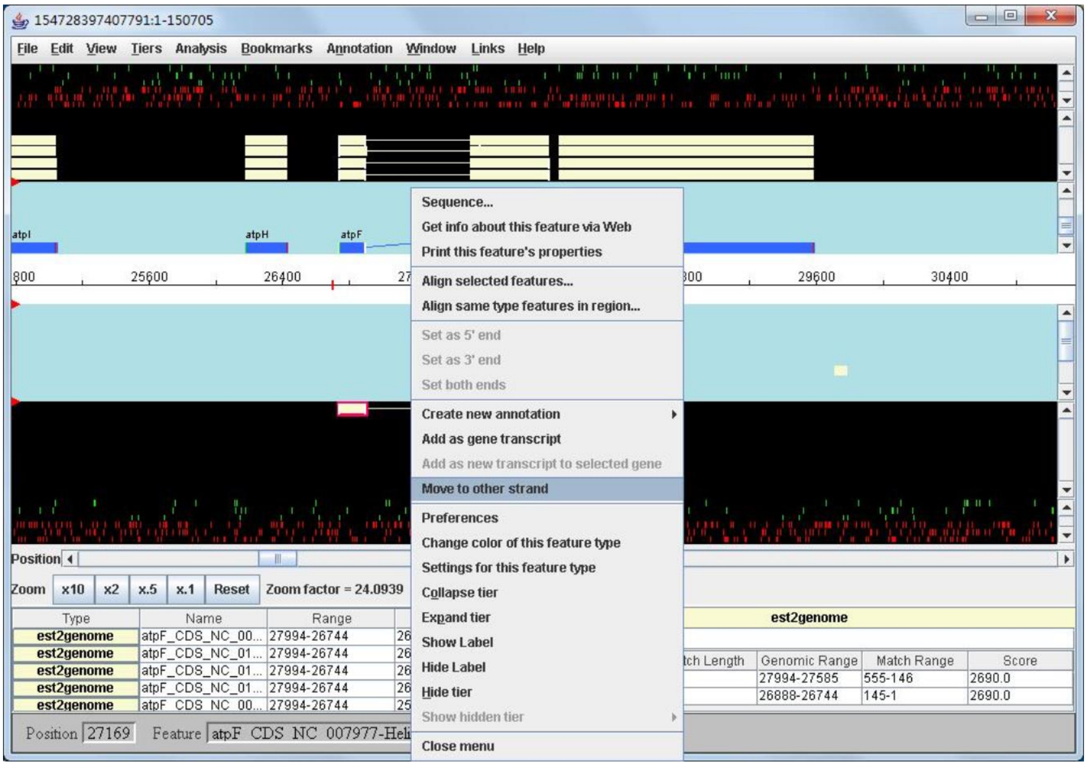
Translation length: 353

☑ Show introns in translation viewer

☐ Follow external selection

IV. Changing the Strand of a Result

If a result feature has been assigned to the wrong strand, you can move it to the other strand. Select the result feature, right click to get to the popup menu, and select "Move to other strand". The result will then be moved to the opposite strand. (If you've selected more than one result, all of them will be moved.)



Type	Name	Range	
est2genome	atpF_CDS_NC_00...	27994-26744	26
est2genome	atpF_CDS_NC_01...	27994-26744	26
est2genome	atpF_CDS_NC_01...	27994-26744	26
est2genome	atpF_CDS_NC_01...	27994-26744	26
est2genome	atpF_CDS_NC_00...	27994-26744	25

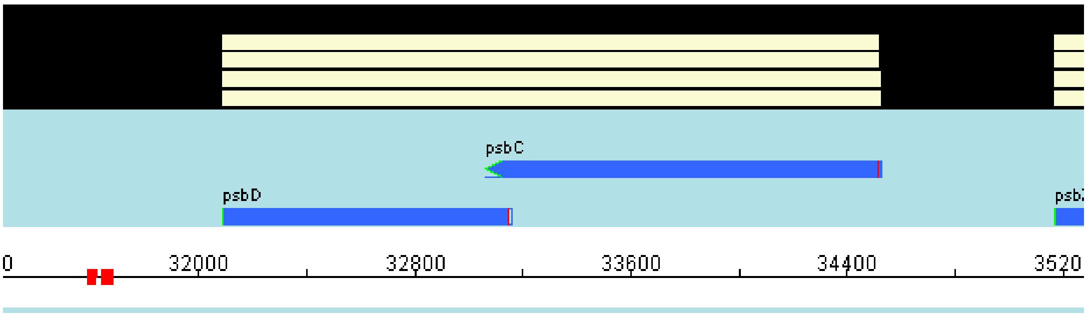
est2genome

tch Length	Genomic Range	Match Range	Score
	27994-27585	555-146	2690.0
	26888-26744	145-1	2690.0

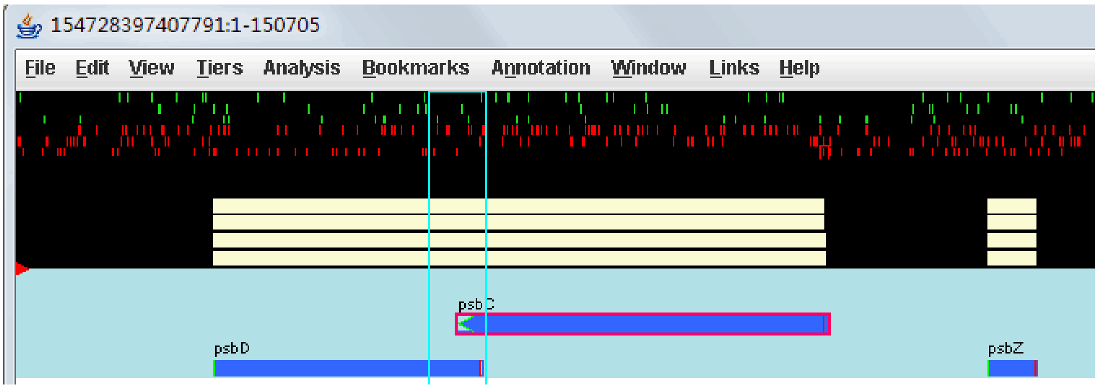
Position 27169 Feature atpF CDS NC 007977-Heli

V. Adjust Start or Stop Codons

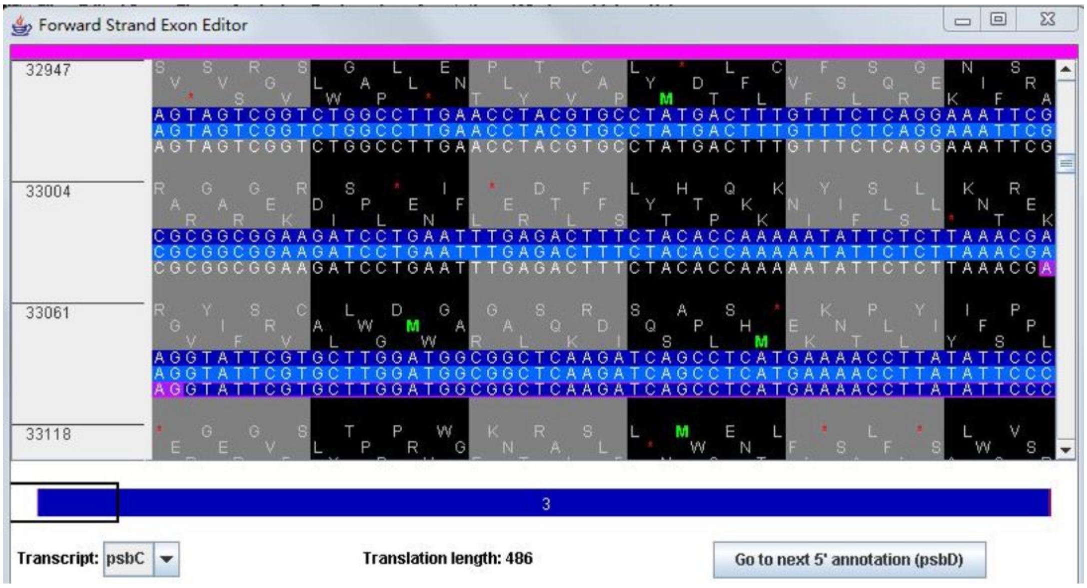
1). If a transcript is missing a start codon, it will appear in the main display with a green arrowhead, and if it is missing a stop codon, it will appear with a red arrowhead. This will also be indicated in the annotation info editor next to "Missing start codon" or "Missing stop codon".



2). Invoke "Exon Detail Editor"



3). Click and drag on the exon to adjust the start codon or stop codon.



```
32947
AGTAGTCGGTCTGGCCTTGAACCTACGTGCCTATGACTTTGTTTCTCAGGAAATTCG
AGTAGTCGGTCTGGCCTTGAACCTACGTGCCTATGACTTTGTTTCTCAGGAAATTCG
AGTAGTCGGTCTGGCCTTGAACCTACGTGCCTATGACTTTGTTTCTCAGGAAATTCG
33004
CGCGGCGGAAGATCCTGAATTTGAGACTTTCTACACCAAAAATATTCTCTTAAACGA
CGCGGCGGAAGATCCTGAATTTGAGACTTTCTACACCAAAAATATTCTCTTAAACGA
CGCGGCGGAAGATCCTGAATTTGAGACTTTCTACACCAAAAATATTCTCTTAAACGA
33061
AGGTATTCGTGCTTGGATGGCGGCTCAAGATCAGCCTCATGAAAACCTTATATTCCC
AGGTATTCGTGCTTGGATGGCGGCTCAAGATCAGCCTCATGAAAACCTTATATTCCC
AGGTATTCGTGCTTGGATGGCGGCTCAAGATCAGCCTCATGAAAACCTTATATTCCC
33118
```

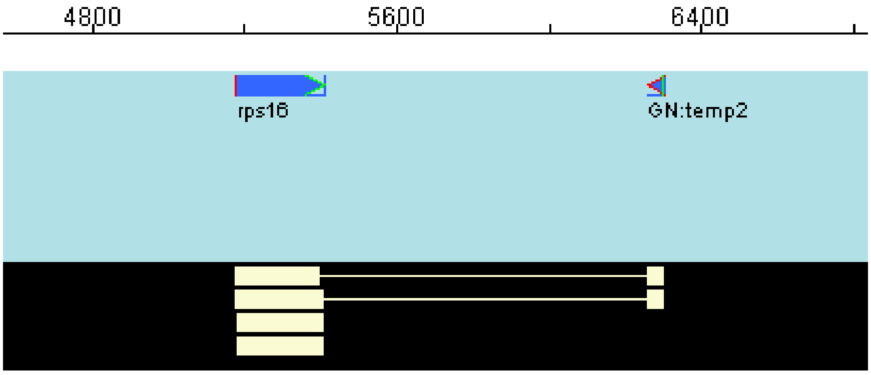
Transcript: psbC

Translation length: 486

VI. Add Missing Exons

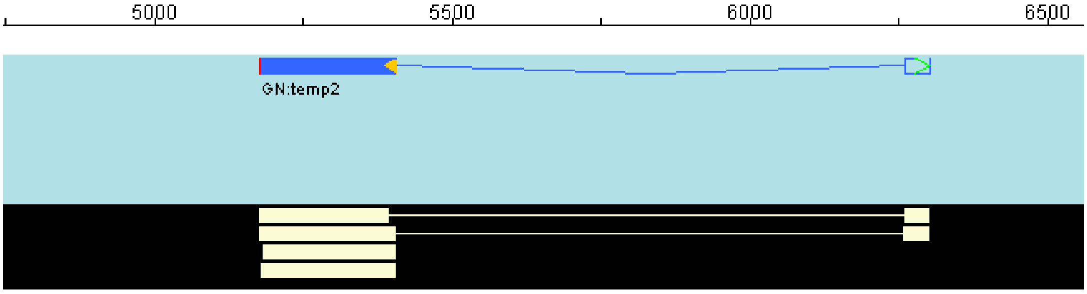
1). Create a new annotation

To create a new annotation, you can simply select results on which to base the annotation and drag them into the light blue annotation zone or use the right mouse popup menu options "Add as gene transcript".



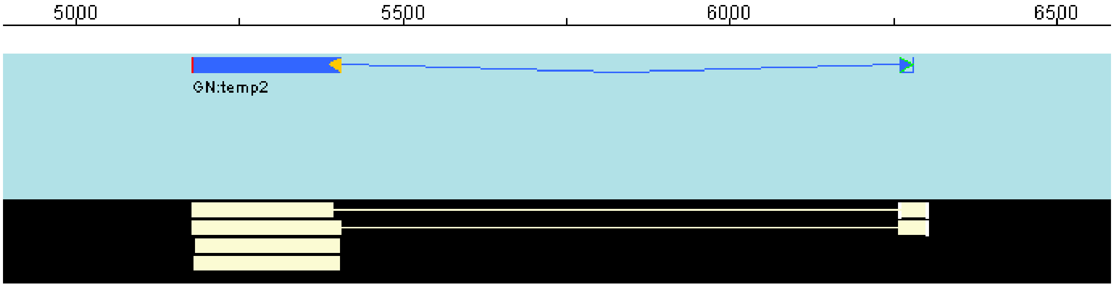
2). Merge transcripts

First, select the transcript A through a single left mouse click. Second, depress the shift key and select the transcript B with a single left mouse click. Third, right click to get the popup menu and select "Merge transcripts".



3). Invoke "Exon Detail Editor" and find appropriate splice site acceptor and donor, respectively.

You can click and drag on the exons to trim the 3' edge of the upstream exon to an appropriate splice site donor, and trim the 5' edge of the downstream exon to an appropriate splice site acceptor.



============================END=================================