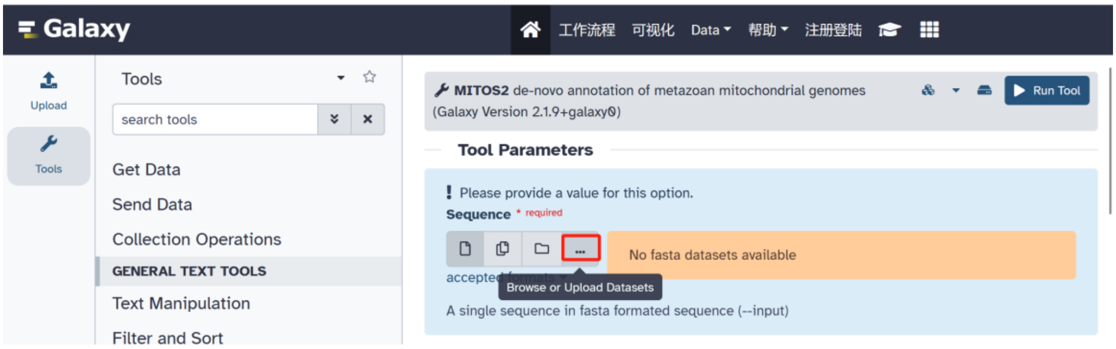
Manual of SFMA
V1.0

Runmao Lin, Tong Liu, Xiaoting Wang, Fanxing Yang, Zhiyin Wang

July, 2023

Step1. Using Mitos web server to predict genes

Upload the mitogenome sequences ("mitogenome.update.v1.fa") to the Mitos web server (https://usegalaxy.org/root?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fiuc%2Fmitos2%2Fmitos2%2F2.1.3%20galaxy0；Reference：RefSeq 63 fungi; Genetic Code: 4 Mold) for gene prediction. The Outputs choose "zipped raw results", which contains one file named "result.gff".



MITOS2 de-novo annotation of metazoan mitochondrial genomes (Galaxy Version 2.1.9+galaxy0)

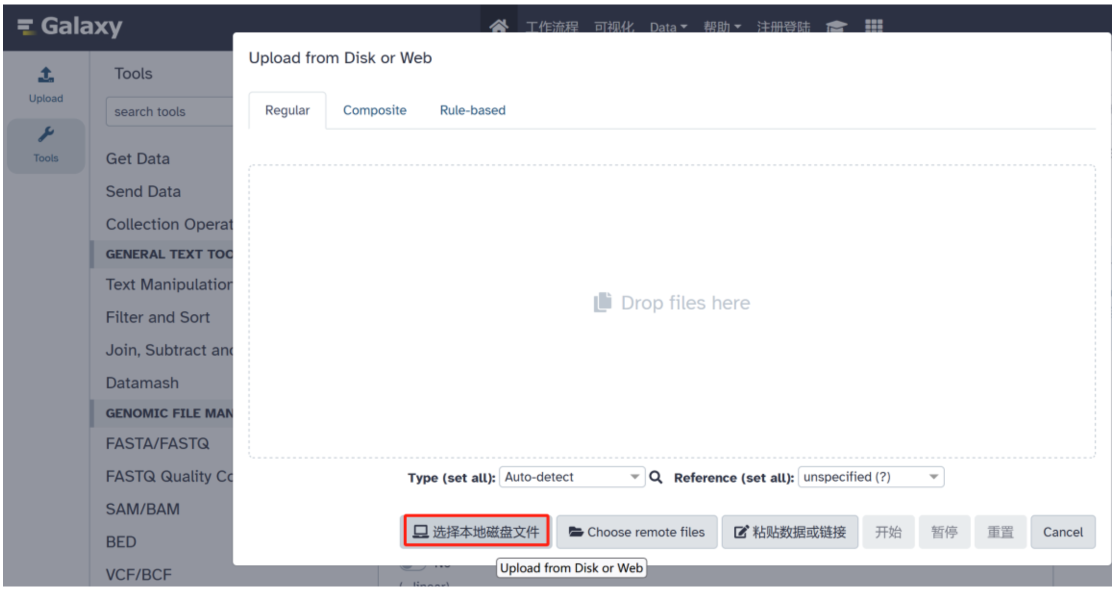
Please provide a value for this option.

Sequence * required

No fasta datasets available

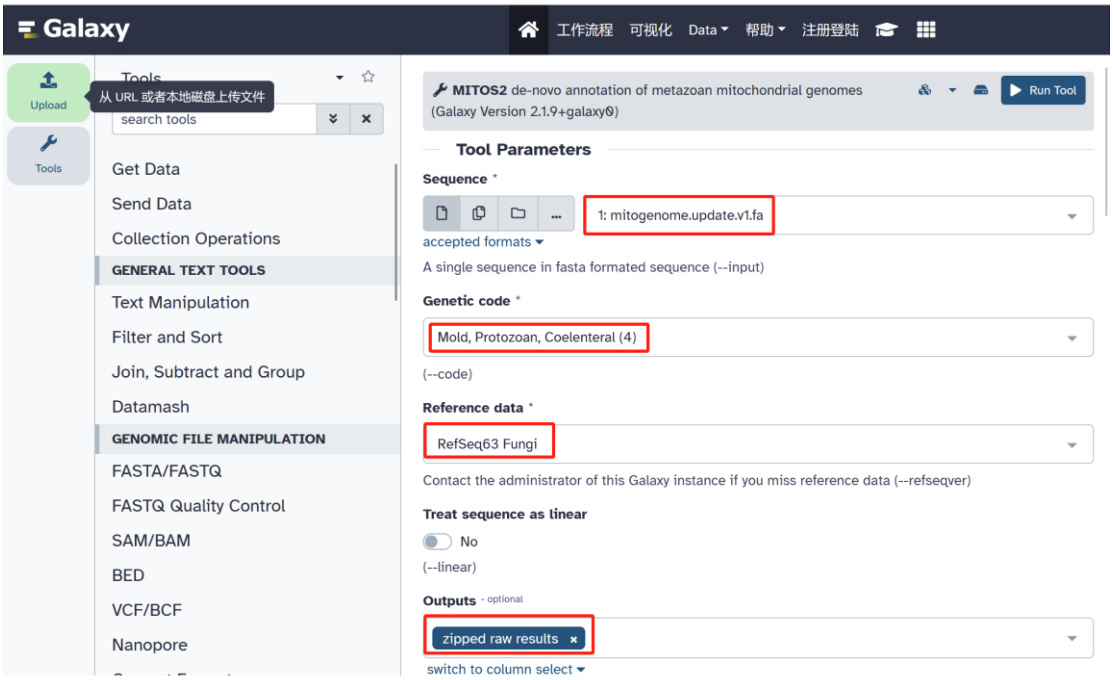
accepted formats

A single sequence in fasta formated sequence (--input)



Type (set all): Auto-detect

Reference (set all): unspecified (?)



MITOS2 de-novo annotation of metazoan mitochondrial genomes (Galaxy Version 2.1.9+galaxy0)

Tool Parameters

Sequence *

1: mitogenome.update.v1.fa

accepted formats

A single sequence in fasta formated sequence (--input)

Genetic code *

Mold, Protozoan, Coelenteral (4)

(--code)

Reference data *

RefSeq63 Fungi

Contact the administrator of this Galaxy instance if you miss reference data (--refseqver)

Treat sequence as linear

No

(--linear)

Outputs - optional

zipped raw results ×

switch to column select

Use "02.mitos_gff_update.v3.pl" to update "result.gff" and generate "mitos_update.gene.gff".

Use "00.fungal_mitogenome_gff2cds_check_complete.v1.pl" to exact nucleotide acids of protein-coding genes.

Use "00.fungal_mitogenome_cds2aa.v1.pl" to translate nucleotide acids to amino acids.

Be careful, if one gene encoded in the "mitogenome.update.v1.fa", the circular mitogenome, was split into two segments at the beginning and end of genome sequences, respectively, users should use "00.split_circular_genome_sequence.v1.pl" to re-split the mitogenome sequence (generated "mitogenome.update.split.v3.fa") and re-annotated using Mitos for analysis.

Examples:

```
perl split_circular_genome_sequence.pl -genome mitogenome.update.v1.fa -position_site position_of_split_site.txt -output mitogenome.update.split.v3.fa
generated mitogenome.update.split.v3.fa
the content of "position_of_split_site.txt":
T203_mitogenome   25069

perl mitos_gff_update.pl -mitos_result_gff result.gff -output_prefix mitos_update
generated mitos_update.gene.gff

perl fungal_mitogenome_gff2cds_check_complete.pl -genome mitogenome.update.v1.fa -gff_file mitos_update.gene.gff -output_file mitos_update.gene.cds
generated mitos_update.gene.cds。

perl fungal_mitogenome_cds2aa.pl -cds_file mitos_update.gene.cds -aa_file mitos_update.gene.pep
```

```
generated mitos_update.gene.pep。
```

Step2. Using FMannot for gene prediction

Upload the mitogenome sequences ("mitogenome.update.v3.fa") to the MFannot web server (https://megasun.bch.umontreal.ca/apps/mfannot/; Genetic Code: 4 Mold) for gene prediction and obtain the "*. fasta.new.tbl" from the downloaded zipped file.

Update "*. fasta.new.tbl" by changing the content in the first line, i.e., change "Feature C_0 Table1" to "Feature T203_mitogenome Table1"; the name of "T203_mitogenome" is the sequence ID of "mitogenome.update.v3.fa".

Use "MFanno_tbl2gff.pl" to update the annotation from MFannot results.

Use "00.fungal_mitogenome_gff2cds_check_complete.v1.pl" to exact nucleotide acids of protein-coding genes.

Use "00.fungal_mitogenome_cds2aa.v1.pl" to translate nucleotide acids to amino acids.

Examples:

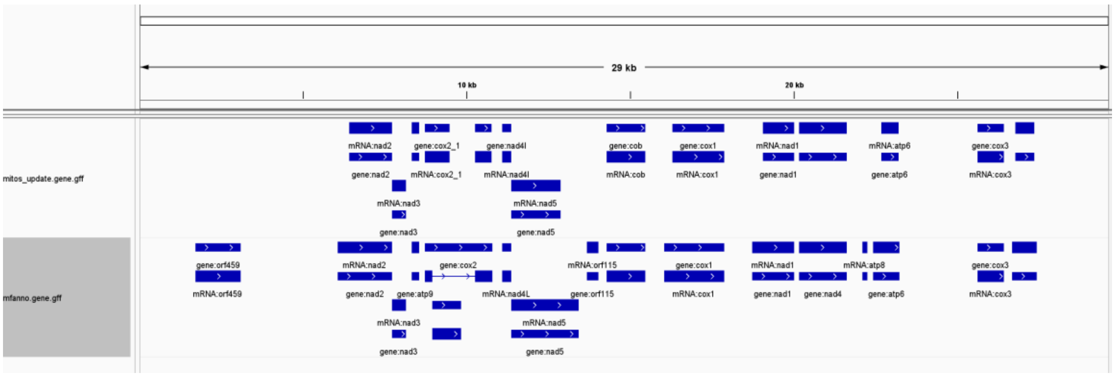
```
perl MFanno_tbl2gff.pl -mfannot_tbl mfannot_1fc54c635917.fasta.new.tbl    -output_prefix mfanno_update
output files include: mfanno.gene.gff, mfanno.rRNA.gff, mfanno.tRNA.gff

perl fungal_mitogenome_gff2cds_check_complete.pl -genome mitogenome.update.v1.fa -gff_file mfanno_update.gene.gff -output_file mfanno_update.gene.cds
output file: mfanno_update.gene.cds

perl fungal_mitogenome_cds2aa.pl -cds_file mfanno_update.gene.cds -aa_file mfanno_update.gene.pep
output file: mfanno_update.gene.pep
```

Step3. Integration of predictions from Mitos and MFannot.

Use "09.comparison_mitos_MFannot.v1.pl" to compare the predicted results from Mitos and MFannot.

```
perl comparison_mitos_MFannot.pl -mitos_update_gff mitos_update.gene.gff -mfanno_update_gff MFannot\mfanno.gene.gff -output_file    infor.txt
obtain file: infor.txt
```

Use IGV (https://igv.org/) to display the results of Mitos and MFannot:

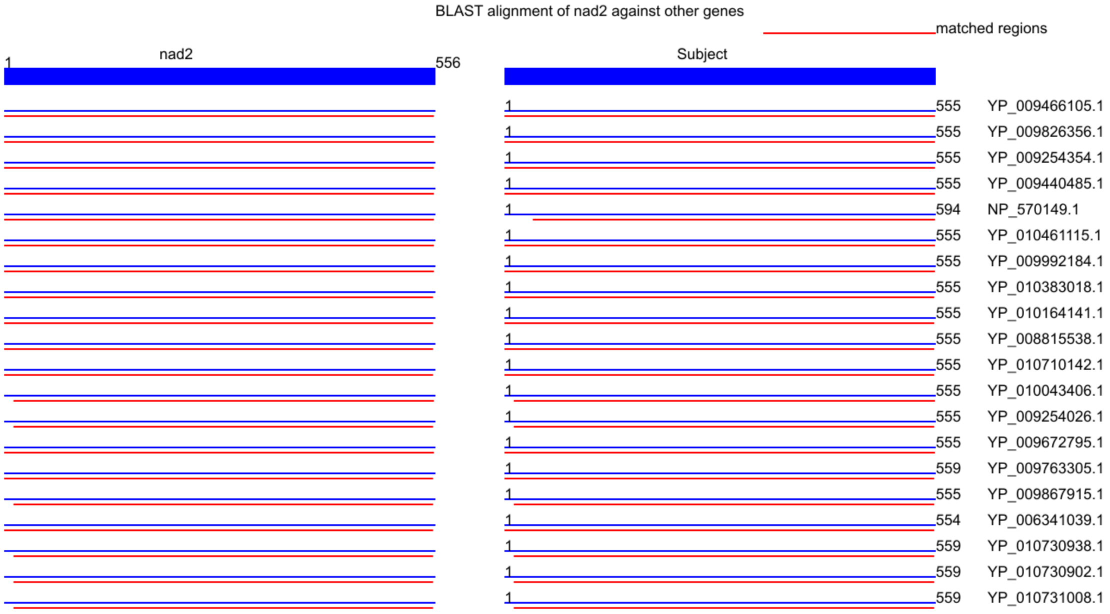


Display of results from Mitos and MFannot by IGV software

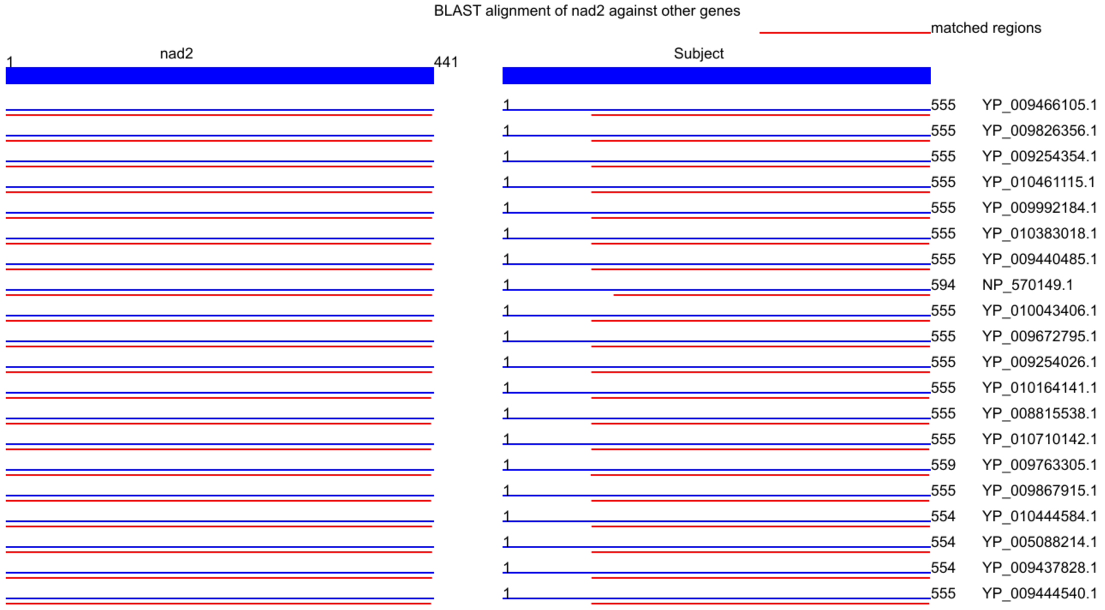
Perform sequence alignment by aligning amino acid sequences of predicted genes from Mitos and MFannot against NCBI refseq mitochondrial genes using BLASTP.

```
D:\03.other\software\blast\blast+\blast-2.14.0+\bin\blastp.exe -query ..\02.MITOS\mitos_update.gene.pep -out mitos2mito.m6.tab -db D:\03.other\software\blast\refseq_mitochondrion\mitochondrion.1.protein.faa -outfmt 6 -evalue 1e-5 -num_threads 2
ouput file: mitos2mito.m6.tab

perl pep_align2refseq_mitochondrion.pl -gene_pep mitos_update.gene.pep -refseq_mitochondrion_pep D:\03.other\software\blast\refseq_mitochondrion\mitochondrion.1.protein.faa -blast_tab mitos2mito.m6.tab -output_prefix mitos
output files, i.e., svg files for viewing alignment of mitos gene sequences

D:\03.other\software\blast\blast+\blast-2.14.0+\bin\blastp.exe -query ..\03.MFannot\mfanno.gene.pep -out mfanno2mito.m6.tab -db D:\03.other\software\blast\refseq_mitochondrion\mitochondrion.1.protein.faa -outfmt 6 -evalue 1e-5 -num_threads 2
ouput file: mfanno2mito.m6.tab

perl pep_align2refseq_mitochondrion.pl -gene_pep mfanno.gene.pep -refseq_mitochondrion_pep D:\03.other\software\blast\refseq_mitochondrion\mitochondrion.1.protein.faa -blast_tab mfanno2mito.m6.tab -output_prefix MFanno
output files, i.e., svg files for viewing alignment of MFannot gene sequences
```

Based on "infor.txt", we found that the predicted nad2 gene from Mitos and MFannot were different. We display the BLAST alignment:



Comparison of alignments between nad2 predicted by MFannot and homoglogous genes (Subject) collected in refseq database



Comparison of alignments between nad2 predicted by Mitos and homoglogous genes (Subject) collected in refseq database

From the alignments as shown above, we found that the nad2 predicted by MFannot may be better, for nad2 predicted by Mitos is shorter that the reported homologous genes collected in refseq. The we selected the MFannot nad2 and wrote it in the "select_gene.txt".

```
The content of "select_gene.txt":
gff_file        GeneID        NewID
```

```
mfanno.gene.gff     atp9    atp9
mfanno.gene.gff     cob     cob
mfanno.gene.gff     cox3    cox3
mfanno.gene.gff     nad3    nad3
mfanno.gene.gff     nad4    nad4
mfanno.gene.gff     nad4L nad4L
mfanno.gene.gff     nad2    nad2
mfanno.gene.gff     cox2    cox2
mfanno.gene.gff     orf294 T203_orf294
mfanno.gene.gff     nad5    nad5
mfanno.gene.gff     orf115 T203_orf115
mfanno.gene.gff     cox1    cox1
mfanno.gene.gff     nad1    nad1
mfanno.gene.gff     atp8    atp8
mfanno.gene.gff     atp6    atp6
mfanno.gene.gff     nad6    nad6
mfanno.gene.gff     orf459 T203_orf459
```

Use "09.select_mitos_MFannot_gff.v1.pl" to select candidate genes from Mitos or MFannot results.

Use "00.fungal_mitogenome_gff2cds_check_complete.v1.pl" to exact nucleotide acids of protein-coding genes.

Use "00.fungal_mitogenome_cds2aa.v1.pl" to translate nucleotide acids to amino acids.

```
perl select_mitos_MFannot_gff.pl -select_gene select_gene.txt -mitos_update_gff mitos_update.gene.gff -mfanno_update_gff mfanno.gene.gff -output_file integrated_gene.gff
output file: integrated_gene.gff

perl fungal_mitogenome_gff2cds_check_complete.pl -genome mitogenome.update.v1.fa -gff_file integrated_gene.gff -output_file integrated_gene.cds
output file: integrated_gene.cds。

perl fungal_mitogenome_cds2aa.pl -cds_file integrated_gene.cds -aa_file integrated_gene.pep
output file: integrated_gene.pep。
```

Step4. Annotation of tRNAs

Upload the mitogenome sequences ("mitogenome.update.v3.fa") to the tRNAscan-SE web server (http://trna.ucsc.edu/tRNAscan-SE/; sequence source: other mitochondrial; Search mode: default; Genetic Code for tRNA Isotype Prediction: Mold & Protozoan Mito) for tRNA annotation.

Then download the Results (i.e., "*.out" file), the Predicted tRNA Secondary Structures (i.e., "*.SS), the Candidate tRNA Sequences in FASTA format (i.e., "*.fa).

Update the "Sequence Name" in the "*.out", with the same as shown in "mitogenome.update.v3.fa".

Use "integrate_tRNA_infor.pl" to integrate the predicted tRNAs from different methods.

```
perl integrate_tRNA_infor.pl -mitos_tRNA_gff mitos_update.tRNA.gff -mfannot_tRNA_gff mfanno.tRNA.gff -tRNAscan_out tRNAscan-SE.seq248545.update.out -output_prefix tRNA
output file: tRNA_result.gff
```

Step5. Annotation of rRNAs

The annotations were mainly from different methods of Mitos, MFannot, Rfam and RNAweasel.

```
The content of rRNA.gff:
T203_mitogenome   mitfi   rRNA  3726   7859   .   +   .   ID=rRNA:rnl;Name=rnl
T203_mitogenome   mitfi   intron 5241  7700   .   +   .   ID=rRNA:rnl;Name=rnl
T203_mitogenome   mitfi   rRNA  28513 29574   .   +   .   ID=rRNA:rns;Name=rns
```

Step6. Annotation of repetitive sequences

Use TRF (https://tandem.bu.edu/trf/trf.html; Submit a Sequence; Basic) for tandem repeat annotation, and obtain "tandem_repeats_finder.summary" and "tandem_repeats_finder.details".

Use "09.mask_exon_seq.v1.pl" to mask gene sequences by Ns and obtain "genome.mask_genes.fa". Then"genome.mask_genes.fa" was used for palindrome annotation (http://palindromes.ibp.cz/#/en/palindrome; Size: 6-30; Spacer: 0-10; Mismatches: 0), we obtain the "DNA_Analyzer_Palindrome.txt".

```
perl mask_exon_seq.pl -genome mitogenome.update.split.v3.fa -gene_gff integrated_gene.gff -tRNA_gff tRNA_result.gff -rRNA_gff rRNA.gff -output_file genome.mask_genes.fa
output file: genome.mask_genes.fa
```

Step7. Draw the circos map

Use " 13.mitogenome_gff_product_ncRNA2tbl.v2.pl" to integrate the annotation results and generate the tbl file.

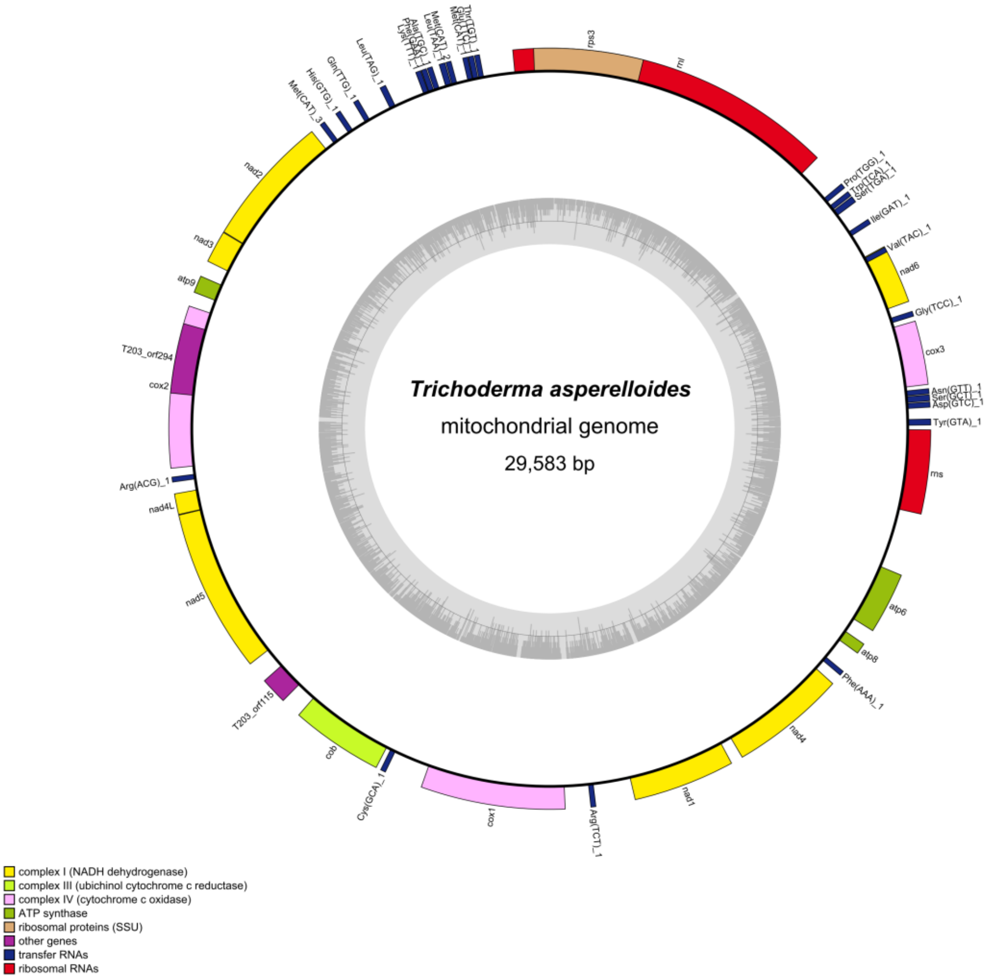
During the analysis, the "gene_product.txt" is required to identify the annotations of genes.

```
The content of "gene_product.txt":
GeneID        ProteinDescription   Domain        PfamAccession
atp9    ATP synthase F0 subunit 9
cob     apocytochrome b
cox3    cytochrome c oxidase subunit 3
nad3    NADH dehydrogenase subunit 3
nad4    NADH dehydrogenase subunit 4
nad4L NADH dehydrogenase subunit 4L
nad2    NADH dehydrogenase subunit 2
cox2    cytochrome c oxidase subunit 2
T203_orf294 GIY-YIG endonuclease
nad5    NADH dehydrogenase subunit 5
T203_orf115 hypothetical protein
cox1    cytochrome c oxidase subunit 1
nad1    NADH dehydrogenase subunit 1
atp8    ATP synthase F0 subunit 8
atp6    ATP synthase F0 subunit 6
nad6    NADH dehydrogenase subunit 6
rps3    ribosomal protein S3
```

Use table2asn to generate the genbank file. And the sbt file can be generated by NCBI server (https://submit.ncbi.nlm.nih.gov/genbank/template/submission/). The mitogenome file was change to the name of "T203.fsa".

Upload the genbank file to OGDRAW (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) and download the svg file of circos map.

```
perl mitogenome_gff_product_ncRNA2tbl.pl -gene_gff integrated_gene.gff -gene_product gene_product.txt -tRNA_gff tRNA_result.gff -rRNA_gff rRNA.gff -dbname HNU -output_file T203.tbl
output file: T203.tbl

table2asn -t T203.sbt -indir ./ -M n -a s -V vb -Z -j "[organism=Trichoderma asperelloides][strain=T203][mgcode=4][location=mitochondrion][topology=circular]"
output file: T203.gbf, i.e., genbank file.
```



The circos map of T203 mitogenome.