[SUPPLEMENTARY]
AGORA : Organellar genome annotation from the amino acid and nucleotide references

Jaehee Jung[1], Jong Im Kim[2], Young-Sik Jeong[3], and Gangman Yi*[3]

[1]Department of General Education, Hongik University, Seoul, Korea
[2]Department of Biology, Chungnam National University, Daejeon, Korea
[3]Department of Multimedia Engineering, Dongguk University, Seoul, Korea

.

1 Data set

1.1 Input

Organellar gene annotation is conducted as follows. We used the raw read file obtained from next-generation sequencing. NGS reads were used for preparation, sequencing, assembly, and annotation. Thus, AGORA requires only one assembled contig of nucleic acid in FASTA format. As shown in Figure S1 , the user can select one FASTA formatted file that can be uploaded. As reference sequences, sequences of amino acids and nucleotides are required. In AGORA, the user-defined references or automatically generated references from NCBI can be uploaded. As shown in Figure S1 , if users use the accession ID, then reference sequences are not required. In the FASTA file of the user-defined reference, RNA genes should contain trn in the description, as this key is used to generate a GenBank file. The available genome types for the organellar gene in the system are chloroplast and mitochondrion. The maximum matched sub gene count per each contig indicates the maximum BLAST matched count.

1.2 Output

Figure S2 is the screen capture of result page. The supported output files are eight types that include the OGDRAW figure. The top four rows of the result table show values for the user-defined options and the remaining outputs are supplementary outputs that can be downloaded. Users can download BLAST sequence alignments of both amino acids and nucleotides form the “Result” of the 8th row. Each query of amino acid gene identifies the most similar positions, where the count of similar positions depends on the maximum matched sub gene count option value. If “maximum matched sub gene count” is set to *N*, the number of matched BLAST results is less than *N*. Other supported files are amino acid and nucleotide database references. If a user-defined file is uploaded, these two files are identical to the uploaded files; however, if the user inserts the accession number, the references of amino acid and nucleotide are generated and displayed. Moreover, the system also provides files in FASTA format that are matched to the database and query of both the amino acid and nucleotide sequences. Separated files of FASTA format grouping by ②s shown in Figure 3(a) are supplied as results, and the nucleotide is also provided. Finally, the system provides the generated GenBank file based on the matched position and its visualization using OGDRAW tools.

*gangman@dongguk.edu, (Corresponding author)

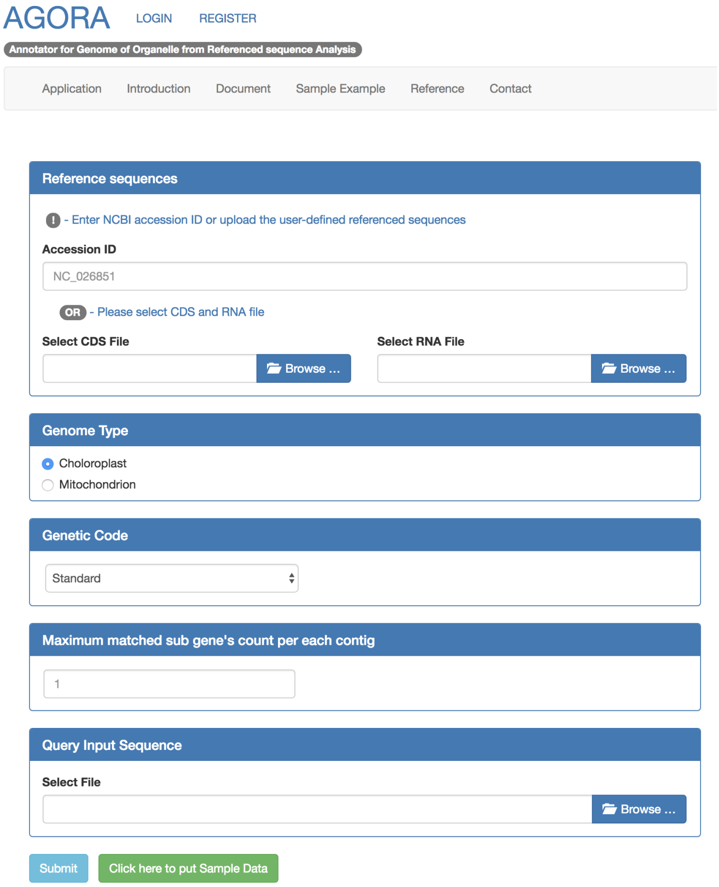
	GeSeq	CpGAVAS	DOGMA	AGORA
Support User define references	✓			✓
Support Chloroplast	✓	✓	✓	✓
Support Mitochondrion	△[a]		△ [a]	✓
Provide BLAST Result				✓
Provide file format	GenBank		GFF3	GenBank
Provide MAP	OGDRAW	GenomeVx	Linear	OGDRAW

Table S1 : Specification comparison of four different systems.

[a] △ stands for not supporting completely.

2 Analysis of results

Table S2 is the comparison analysis of different applications using representative species. On the third column, PL stands for chloroplast and MT is mitochondrion. The fourth column shows the original gene count from the NCBI GenBank file. From the fifth to eighth column, we show the number of genes from four different applications. AGORA found exactly the same number of original genes, but GeSeq resulted in more genes. The reason of different number between AGORA and GeSeq is the fragment of genes. In order to analyze the GenBank file, OGDRAW is employed since this tool is used to draw the circular map by the GenBank file. As Table S3 is displayed, three different OGDRAW images can be compared. The data is randomly selected from the Table S2 . However, CpGAVAS does not provide the mitochondrion, so N/A is represented in the table. DOGMA does not provide as many genes as AGORA and GeSeq either.



Reference sequences

- Enter NCBI accession ID or upload the user-defined referenced sequences

Accession ID: NC_026851

OR - Please select CDS and RNA file

Select CDS File

Select RNA File

Genome Type

● Choloroplast

○ Mitochondrion

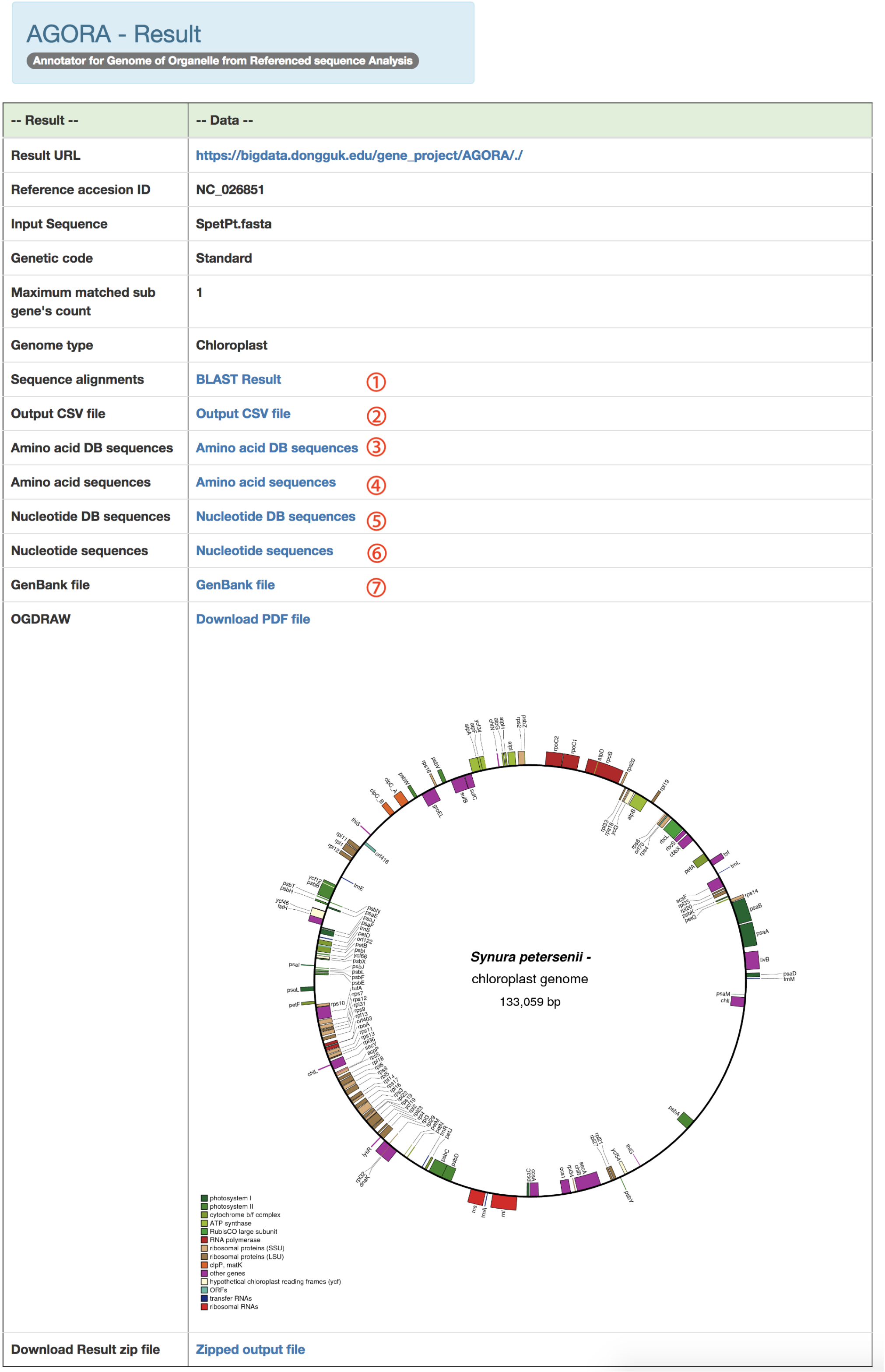
Genetic Code: Standard

Maximum matched sub gene's count per each contig: 1

Query Input Sequence

Select File

Figure S1 : Screen capture of user-defined option values.

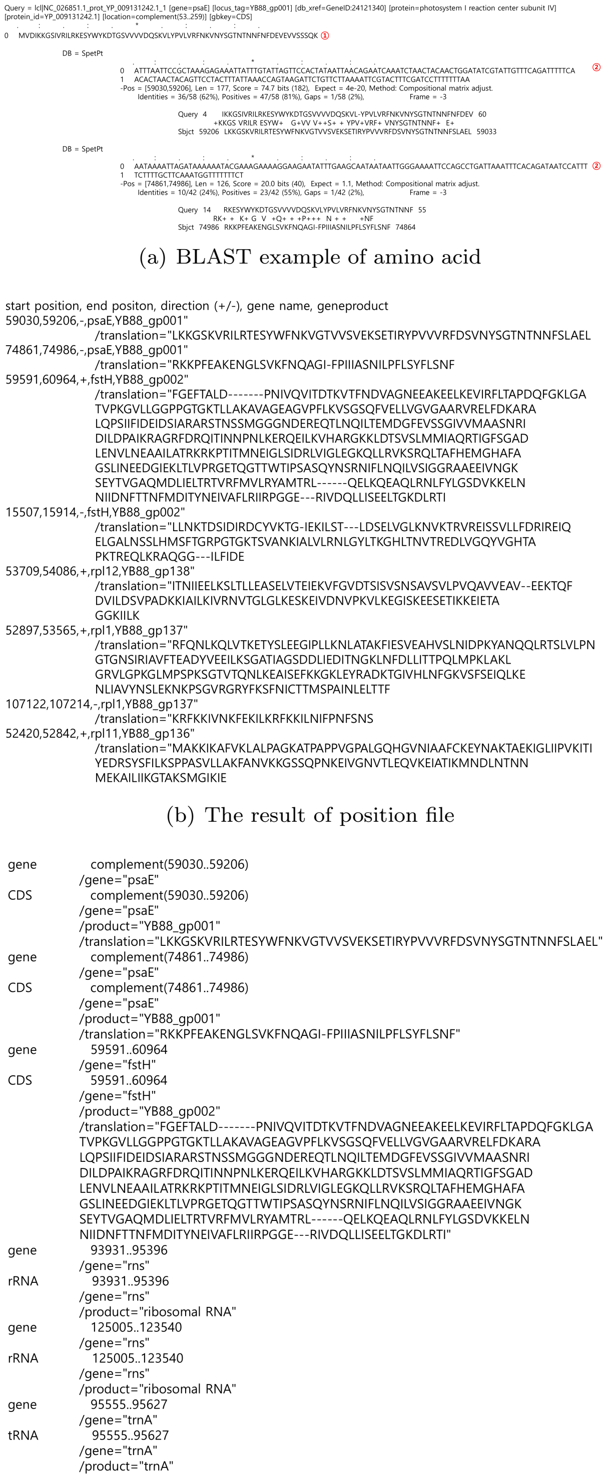


AGORA - Result

Annotator for Genome of Organelle from Referenced sequence Analysis

-- Result --	-- Data --
Result URL	https://bigdata.dongguk.edu/gene_project/AGORA/./
Reference accesion ID	NC_026851
Input Sequence	SpetPt.fasta
Genetic code	Standard
Maximum matched sub gene's count	1
Genome type	Chloroplast
Sequence alignments	BLAST Result ①
Output CSV file	Output CSV file ②
Amino acid DB sequences	Amino acid DB sequences ③
Amino acid sequences	Amino acid sequences ④
Nucleotide DB sequences	Nucleotide DB sequences ⑤
Nucleotide sequences	Nucleotide sequences ⑥
GenBank file	GenBank file ⑦
OGDRAW	Download PDF file
Download Result zip file	Zipped output file

Figure S2 : Screen capture of the AGORA output page. Supplementary files such as BLAST result, matched position file, and GenBank file are provided. ① can be downloaded as BLAST results, with the screen capture shown in Figure S3(a). ② is linked to the separated position file, which includes the start position, end position, direction and product name of each gene. Please refer to Figure S3(b). ③ and ⑤ are sequences of the original reference amino acid and nucleotide. ④ and ⑥ are sets of aligned sequences, of which the screen is shown in Figure S3(a). ⑦ is the GenBank format file shown in Figure S3(c).



```
Query = lcl|NC_026851.1_prot_YP_009131242.1_1 [gene=psaE] [locus_tag=YB88_gp001] [db_xref=GeneID:24121340] [protein=photosystem I reaction center subunit IV]
[protein_id=YP_009131242.1] [location=complement(53..259)] [gbkey=CDS]
    .       :       .       :       .       *       .       :       .       :       .
0    MVDIKKGSIVRILRKESYWYKDTGSVVVVDQSKVLYPVLVRFNKVNYSGTNTNNFNFDEVEVVSSSQK ①

                    DB = SpetPt
                                     .       :       .       :       .       *       .       :       .       :       .
                                  0    ATTTAATTCCGCTAAAGAGAAATTATTTGTATTAGTTCCACTATAATTAACAGAATCAAATCTAACTACAACTGGATATCGTATTGTTTCAGATTTTTCA      ②
                                  1    ACACTAACTACAGTTCCTACTTTATTAAACCAGTAAGATTCTGTTCTTAAAATTCGTACTTTCGATCCTTTTTTTAA
                                  -Pos = [59030,59206], Len = 177, Score = 74.7 bits (182),  Expect = 4e-20, Method: Compositional matrix adjust.
                                        Identities = 36/58 (62%), Positives = 47/58 (81%), Gaps = 1/58 (2%),                  Frame = -3

                                                 Query  4      IKKGSIVRILRKESYWYKDTGSVVVVDQSKVL-YPVLVRFNKVNYSGTNTNNFNFDEV  60
                                                            +KKGS VRILR ESYW+   G+VV V++S+ + YPV+VRF+ VNYSGTNTNNF+  E+
                                                 Sbjct  59206  LKKGSKVRILRTESYWFNKVGTVVSVEKSETIRYPVVVRFDSVNYSGTNTNNFSLAEL  59033

                    DB = SpetPt
                                     .       :       .       :       .       *       .       :       .       :       .
                                  0    AATAAAATTAGATAAAAAATACGAAAGAAAAGGAAGAATATTTGAAGCAATAATAATTGGGAAAATTCCAGCCTGATTAAATTTCACAGATAATCCATTT ②
                                  1    TCTTTTGCTTCAAATGGTTTTTTTCT
                                  -Pos = [74861,74986], Len = 126, Score = 20.0 bits (40),  Expect = 1.1, Method: Compositional matrix adjust.
                                        Identities = 10/42 (24%), Positives = 23/42 (55%), Gaps = 1/42 (2%),                  Frame = -3

                                                 Query  14     RKESYWYKDTGSVVVVDQSKVLYPVLVRFNKVNYSGTNTNNF  55
                                                            RK+ +  K+ G  V  +Q+ + +P+++  N + +     +NF
                                                 Sbjct  74986  RKKPFEAKENGLSVKFNQAGI-FPIIIASNILPFLSYFLSNF  74864
```

(a) BLAST example of amino acid

```
start position, end positon, direction (+/-), gene name, geneproduct
59030,59206,-,psaE,YB88_gp001"
                /translation="LKKGSKVRILRTESYWFNKVGTVVSVEKSETIRYPVVVRFDSVNYSGTNTNNFSLAEL
74861,74986,-,psaE,YB88_gp001"
                /translation="RKKPFEAKENGLSVKFNQAGI-FPIIIASNILPFLSYFLSNF
59591,60964,+,fstH,YB88_gp002"
                /translation="FGEFTALD-------PNIVQVITDTKVTFNDVAGNEEAKEELKEVIRFLTAPDQFGKLGA
                TVPKGVLLGGPPGTGKTLLAKAVAGEAGVPFLKVSGSQFVELLVGVGAARVRELFDKARA
                LQPSIIFIDEIDSIARARSTNSSMGGGNDEREQTLNQILTEMDGFEVSSGIVVMAASNRI
                DILDPAIKRAGRFDRQITINNPNLKERQEILKVHARGKKLDTSVSLMMIAQRTIGFSGAD
                LENVLNEAAILATRKRKPTITMNEIGLSIDRLVIGLEGKQLLRVKSRQLTAFHEMGHAFA
                GSLINEEDGIEKLTLVPRGETQGTTWTIPSASQYNSRNIFLNQILVSIGGRAAEEIVNGK
                SEYTVGAQMDLIELTRTVRFMVLRYAMTRL------QELKQEAQLRNLFYLGSDVKKELN
                NIIDNFTTNFMDITYNEIVAFLRIIRPGGE---RIVDQLLISEELTGKDLRTI
15507,15914,-,fstH,YB88_gp002"
                /translation="LLNKTDSIDIRDCYVKTG-IEKILST---LDSELVGLKNVKTRVREISSVLLFDRIREIQ
                ELGALNSSLHMSFTGRPGTGKTSVANKIALVLRNLGYLTKGHLTNVTREDLVGQYVGHTA
                PKTREQLKRAQGG---ILFIDE
53709,54086,+,rpl12,YB88_gp138"
                /translation="ITNIIEELKSLTLLEASELVTEIEKVFGVDTSISVSNSAVSVLPVQAVVEAV--EEKTQF
                DVILDSVPADKKIAILKIVRNVTGLGLKESKEIVDNVPKVLKEGISKEESETIKKEIETA
                GGKIILK
52897,53565,+,rpl1,YB88_gp137"
                /translation="RFQNLKQLVTKETYSLEEGIPLLKNLATAKFIESVEAHVSLNIDPKYANQQLRTSLVLPN
                GTGNSIRIAVFTEADYVEEILKSGATIAGSDDLIEDITNGKLNFDLLITTPQLMPKLAKL
                GRVLGPKGLMPSPKSGTVTQNLKEAISEFKKGKLEYRADKTGIVHLNFGKVSFSEIQLKE
                NLIAVYNSLEKNKPSGVRGRYFKSFNICTTMSPAINLELTTF
107122,107214,-,rpl1,YB88_gp137"
                /translation="KRFKKIVNKFEKILKRFKKILNIFPNFSNS
52420,52842,+,rpl11,YB88_gp136"
                /translation="MAKKIKAFVKLALPAGKATPAPPVGPALGQHGVNIAAFCKEYNAKTAEKIGLIIPVKITI
                YEDRSYSFILKSPPASVLLAKFANVKKGSSQPNKEIVGNVTLEQVKEIATIKMNDLNTNN
                MEKAILIIKGTAKSMGIKIE
```

(b) The result of position file

```
gene            complement(59030..59206)
              /gene="psaE"
CDS             complement(59030..59206)
              /gene="psaE"
              /product="YB88_gp001"
              /translation="LKKGSKVRILRTESYWFNKVGTVVSVEKSETIRYPVVVRFDSVNYSGTNTNNFSLAEL"
gene            complement(74861..74986)
              /gene="psaE"
CDS             complement(74861..74986)
              /gene="psaE"
              /product="YB88_gp001"
              /translation="RKKPFEAKENGLSVKFNQAGI-FPIIIASNILPFLSYFLSNF"
gene            59591..60964
              /gene="fstH"
CDS             59591..60964
              /gene="fstH"
              /product="YB88_gp002"
              /translation="FGEFTALD-------PNIVQVITDTKVTFNDVAGNEEAKEELKEVIRFLTAPDQFGKLGA
              TVPKGVLLGGPPGTGKTLLAKAVAGEAGVPFLKVSGSQFVELLVGVGAARVRELFDKARA
              LQPSIIFIDEIDSIARARSTNSSMGGGNDEREQTLNQILTEMDGFEVSSGIVVMAASNRI
              DILDPAIKRAGRFDRQITINNPNLKERQEILKVHARGKKLDTSVSLMMIAQRTIGFSGAD
              LENVLNEAAILATRKRKPTITMNEIGLSIDRLVIGLEGKQLLRVKSRQLTAFHEMGHAFA
              GSLINEEDGIEKLTLVPRGETQGTTWTIPSASQYNSRNIFLNQILVSIGGRAAEEIVNGK
              SEYTVGAQMDLIELTRTVRFMVLRYAMTRL------QELKQEAQLRNLFYLGSDVKKELN
              NIIDNFTTNFMDITYNEIVAFLRIIRPGGE---RIVDQLLISEELTGKDLRTI"
gene            93931..95396
              /gene="rns"
rRNA            93931..95396
              /gene="rns"
              /product="ribosomal RNA"
gene            125005..123540
              /gene="rns"
rRNA            125005..123540
              /gene="rns"
              /product="ribosomal RNA"
gene            95555..95627
              /gene="trnA"
tRNA            95555..95627
              /gene="trnA"
              /product="trnA"
```

(c) The results of GenBank format

Figure S3 : Screen capture of three representative results. Figure S3(a) is one of BLAST examples when users set 2 as the "Maximum matched sub gene's count per each contig" at the input option. The query and database for running BLAST are reversed, when the best matched position of the targeted genes is identified. ① in Figure S3(a) is an example of the amino acid result. ② in Figure S3(a) will be used to generate the output of "Amino acid sequences" at shown in Figure S2 ④. Figure S3(b) is the csv file that represents the direction and position of the genes. Figure S3(c) is the input file of OGDRAW showing the final results.

	Species	NCBI RefSeq	Genome Type	Genes Count	AGORA	GeSeq	CpGAVAS	DOGMA
Viridiplantae	Arabidopsis thaliana	NC_000932	PL	129	129	142	139	155
		NC_001284	MT	131	141	208	39	4
Green algae	Chlorella variabilis	NC_015359	PL	115	115	120	107	115
		NC_025413	MT	62	62	73	28	3
Rhodophyta	Gracilaria chorda	NC_031149	PL	233	233	250	N/A	220
		NC_023251	MT	51	51	51	N/A	3
Phaeophyceae	Ectocarpus silicolosus	NC_013498	PL	185	185	212	160	115
		NC_030223	MT	68	68	68	25	2
Anoebozoa	Acanthamoeba castellanii	NC_001637	MT	57	57	57	N/A	1
Opithokonta (Fungi)	Saccharomyces cerevisiae	NC_027264	MT	35	34	39	N/A	15
Human	Homo sapiens	NC_012920	MT	37	37	37	N/A	15
Beetles (Hoeny bee)	Apis mellifera	NC_001566	MT	13	29	37	N/A	15
Mouse	Mus musculus (C57BL/6J)	NC_005089	MT	37	37	37	N/A	15
Fish (Zebra)	Danio rerio	NC_002333	MT	37	37	37	N/A	15

Table S2 : Comparison analysis of different applications using representative species. At the genome type, PL indicates chloroplast and MT is mitochondria. The genes count is the number of genes for the NCBI reference.

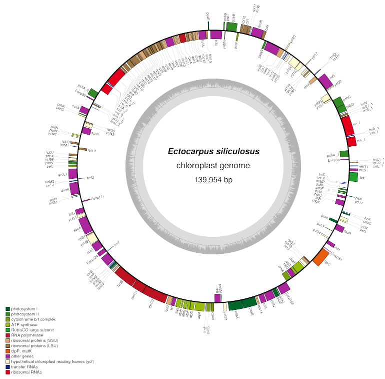
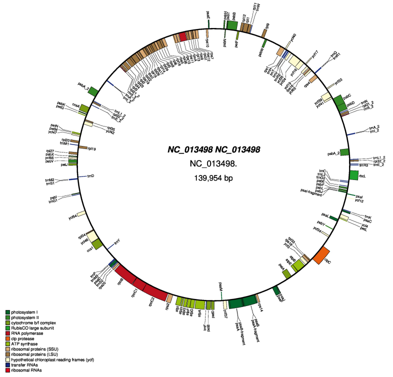
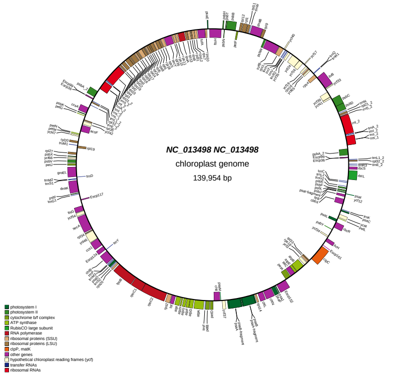
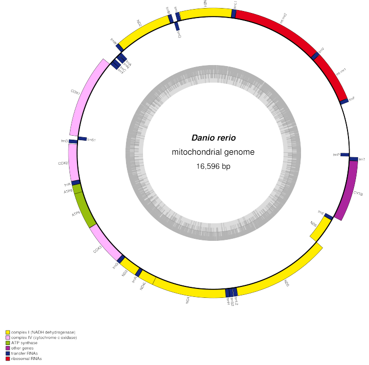
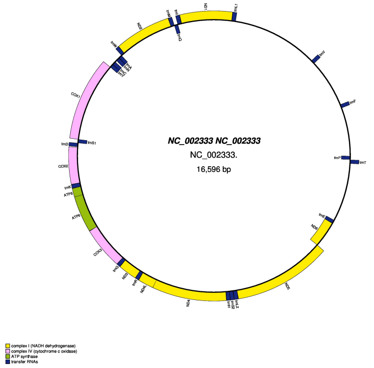
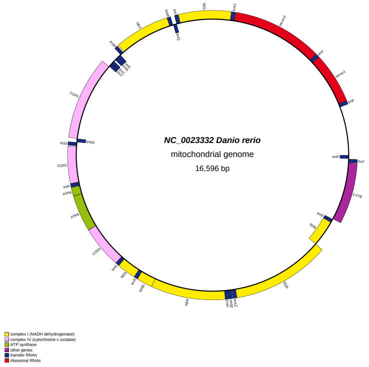
Accession Number	Genome Type	GenBank	GeSeq	AGORA
NC_013498	PL	 *Ectocarpus siliculosus* chloroplast genome 139,954 bp photosystem I photosystem II cytochrome b/f complex ATP synthase RubisCO large subunit RNA polymerase ribosomal proteins (SSU) ribosomal proteins (LSU) clpP, matK other genes hypothetical chloroplast reading frames (ycf) transfer RNAs ribosomal RNAs	 *NC_013498 NC_013498* NC_013498. 139,954 bp photosystem I photosystem II cytochrome b/f complex RubisCO large subunit RNA polymerase clp protease ATP synthase ribosomal proteins (SSU) ribosomal proteins (LSU) hypothetical chloroplast reading frames (ycf) transfer RNAs ribosomal RNAs	 *NC_013498 NC_013498* chloroplast genome 139,954 bp photosystem I photosystem II cytochrome b/f complex ATP synthase RubisCO large subunit RNA polymerase ribosomal proteins (SSU) ribosomal proteins (LSU) clpP, matK other genes hypothetical chloroplast reading frames (ycf) transfer RNAs ribosomal RNAs
NC_002333	MT	 *Danio rerio* mitochondrial genome 16,596 bp complex I (NADH dehydrogenase) complex IV (cytochrome c oxidase) ATP synthase other genes transfer RNAs ribosomal RNAs	 *NC_002333 NC_002333* NC_002333. 16,596 bp complex I (NADH dehydrogenase) complex IV (cytochrome c oxidase) ATP synthase transfer RNAs	 *NC_0023332 Danio rerio* mitochondrial genome 16,596 bp complex I (NADH dehydrogenase) complex IV (cytochrome c oxidase) ATP synthase other genes transfer RNAs ribosomal RNAs

Table S3 : Comparison analysis of OGDRAW. The third column is drawn from the original Genbank file, the fourth column is the image of GeSeq image and fifth column is drawn by AGORA.