

BioAider V1.03

1. Introduction

2. Download and install

3. Functions

3.1. SeqTools

3.1.1. Seqformat Convertor

3.1.2. SeqVary

3.1.3. SequenceID Rename

3.1.4. Split Sequence Fragment

3.1.5. Combine Gene (Tandem Gene)

3.1.6. Visual Gene Extractor

3.1.7. Fast Annotation

3.2. Similar Analysis

3.2.1. Sequence Identity Matrix

3.2.2 Remove H-Similar Sequence

3.3. Align tools

3.3.1. Mafft

3.3.2. Muscle

3.3.3. Clustal-Omeg

3.4. Mutation Tools

3.4.1. Mutation Analysis

3.4.2. Site Counter

3.4.3. Site Scree

4. Test Datas

Manual of BioAider V1.03

A richly featured desktop platform libraries for
analysis of bioinformatics datas

Home page: <https://github.com/ZhijianZhou01/BioAider>

Version 1.03 || August 18, 2020

1. Introduction

With the development of sequencing technology, a large amount of genomic sequenced data has been accumulated. Analyzing these data will help us understand their genetic variation at the molecular level. However, processing a large-scale sequences is difficult for biological or clinical expert without bioinformatics and programming skills. Besides, the needs are also diverse due to different research purposes. Therefore, simplicity of operation and diversity of function are needed.

Bioinformatics Aider (BioAider) V1.03 is developed based on Python3 and PySide2, which is a user-friendly GUI-interface program. As a desktop platform for genomic sequencing data studies, BioAider is designed to simplicity of operation and high summary of analysis results, which could save a lot of time for researchers.

2. Download and install

BioAider and all the updated versions is freely available for non-commercial user at <https://github.com/ZhijianZhou01/BioAider/releases>. After obtaining the program, users could directly run the program by clicking executable file in Windows or Linux(Ubuntu 16.04 or more) systems without installation.

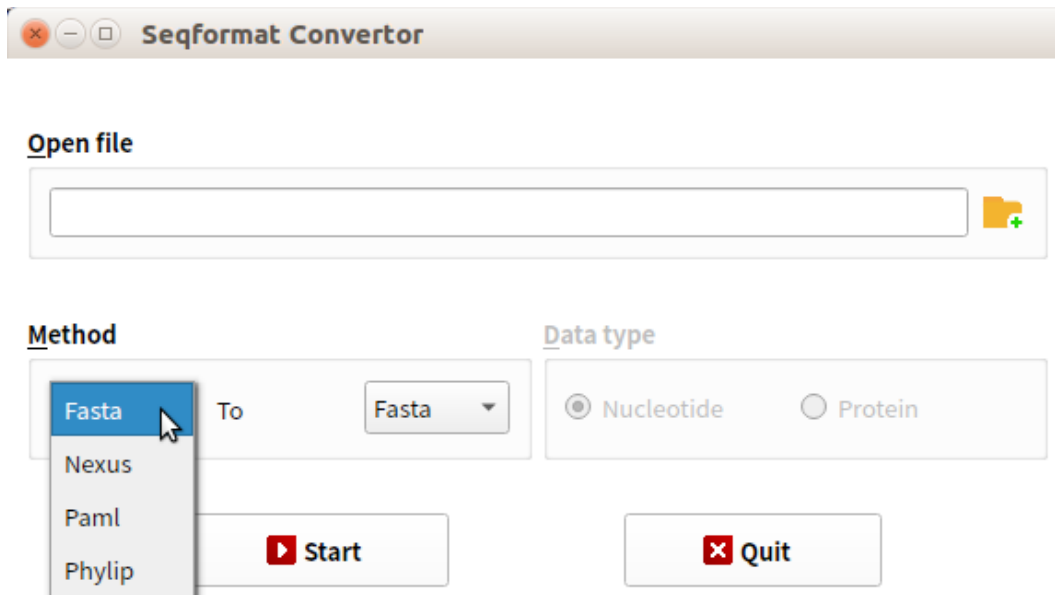
BioAider will be in the long-term update, this document briefly introduces some of its current commonly functions.

3. Functions

3.1. *SeqTools*

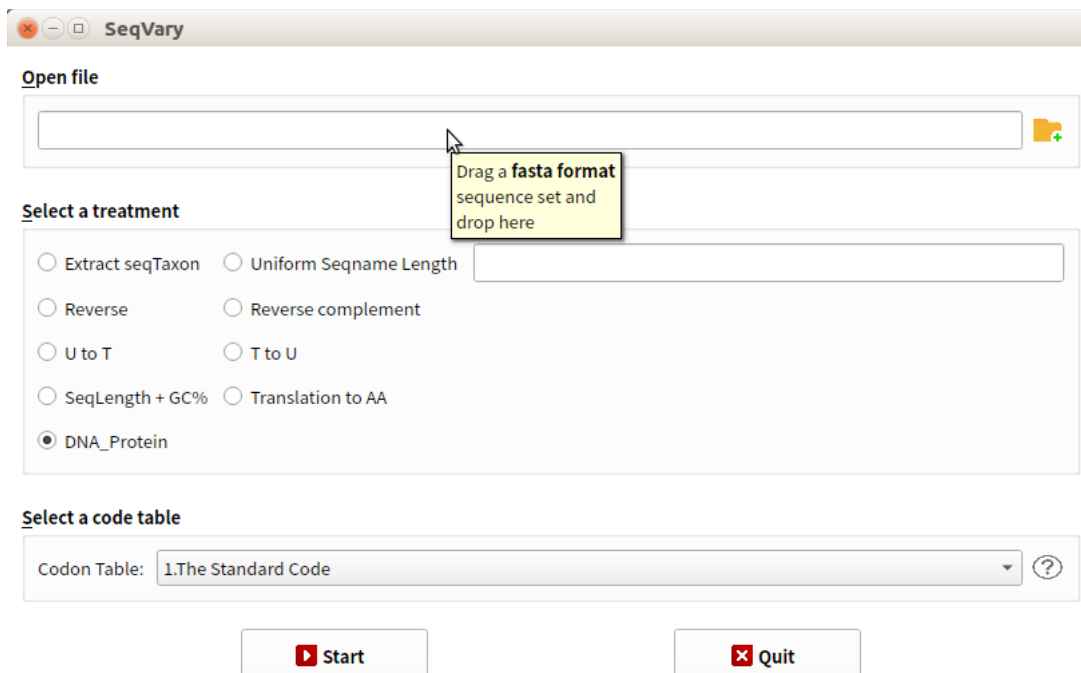
3.1.1. *Seqformat Convertor*

BioAider provides mutual conversion among several common sequence formats, which are Fasta, Nexus, Paml, and Phylip. Of note, the "*Data type*" option is only available when the target format is "Nexus".



3.1.2. SeqVary

The ***"SeqVary"*** option of BioAider provides some small functions for sequence preprocessing. For example, ***"SeqLength+GC%"*** is used to batch calculate sequence length and content of GC.



Note: the ***"DNA_Protein"*** option requires the gene sequences datas to be aligned based on codons.

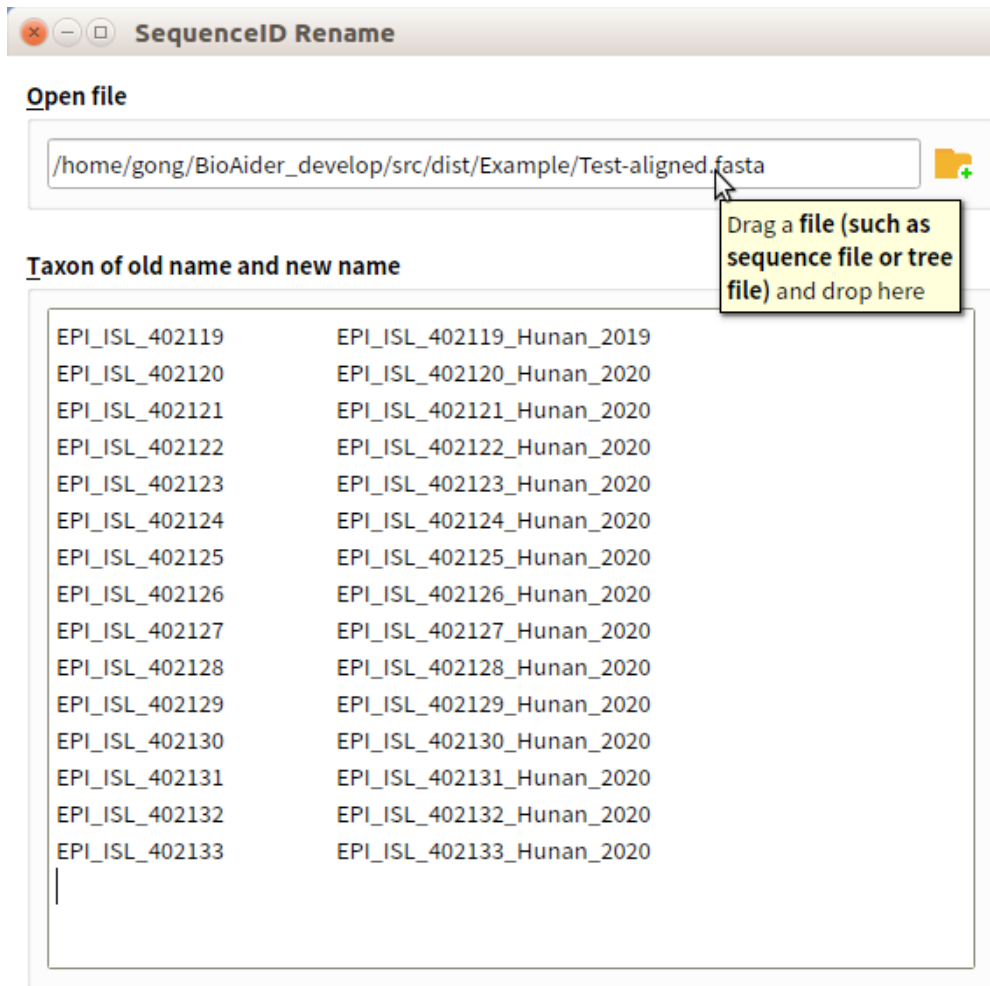
3.1.3. *SequenceID Rename*

BioAider could rename the original name in **sequence datas or tree file etc.** In particular, the pictures of the evolutionary tree used for publication often require the taxons of tree to follow a uniform format, so first batch replacement in the tree file saves the trouble of using vector graphics tools to modify later.

First, make a table of **new and old names** in a table editor :

A2:B16	f _x	Σ	=	EPI_ISL_402119
	A	B	C	D
1	Old name	New name		
2	EPI_ISL_402119	EPI_ISL_402119_Hunan_2019		
3	EPI_ISL_402120	EPI_ISL_402120_Hunan_2020		
4	EPI_ISL_402121	EPI_ISL_402121_Hunan_2020		
5	EPI_ISL_402122	EPI_ISL_402122_Hunan_2020		
6	EPI_ISL_402123	EPI_ISL_402123_Hunan_2020		
7	EPI_ISL_402124	EPI_ISL_402124_Hunan_2020		
8	EPI_ISL_402125	EPI_ISL_402125_Hunan_2020		
9	EPI_ISL_402126	EPI_ISL_402126_Hunan_2020		
10	EPI_ISL_402127	EPI_ISL_402127_Hunan_2020		
11	EPI_ISL_402128	EPI_ISL_402128_Hunan_2020		
12	EPI_ISL_402129	EPI_ISL_402129_Hunan_2020		
13	EPI_ISL_402130	EPI_ISL_402130_Hunan_2020		
14	EPI_ISL_402131	EPI_ISL_402131_Hunan_2020		
15	EPI_ISL_402132	EPI_ISL_402132_Hunan_2020		
16	EPI_ISL_402133	EPI_ISL_402133_Hunan_2020		
17				

Then copy and paste them into BioAider:



Generally speaking, as long as the input file is a text file, BioAider could successfully perform this work.

3.1.4. *Split Sequence Fragment*

This function can batch intercept the specified range of gene fragments, two different modes are available: specified different range ("***Different range***") for each sequence, equal range for all sequences ("***Equal range***").

If you want to use the "***Different range***" to split for each sequence , make a table of start and end location firstly:

A2:C16				
fx Σ = EPI_ISL_402119				
	A	B	C	D
1	Name	Start	End	
2	EPI_ISL_402119	1	600	
3	EPI_ISL_402120	3	700	
4	EPI_ISL_402121	46	788	
5	EPI_ISL_402122	7	888	
6	EPI_ISL_402123	9	333	
7	EPI_ISL_402124	5	888	
8	EPI_ISL_402125	33	777	
9	EPI_ISL_402126	23	679	
10	EPI_ISL_402127	33	767	
11	EPI_ISL_402128	33	767	
12	EPI_ISL_402129	55	890	
13	EPI_ISL_402130	33	900	
14	EPI_ISL_402131	44	678	
15	EPI_ISL_402132	38	876	
16	EPI_ISL_402133	56	890	
17				
18				
19				

Then copy and paste them into BioAider:

Split Sequence Fragment

Open file

/home/gong/BioAider_develop/src/dist/Example/Test-aligned.fasta

Mode of Split

☒ Different range
☐ Equal range

Taxon with seqID and regional

EPI_ISL_402119	1	600
EPI_ISL_402120	3	700
EPI_ISL_402121	46	788
EPI_ISL_402122	7	888
EPI_ISL_402123	9	333
EPI_ISL_402124	5	888
EPI_ISL_402125	33	777
EPI_ISL_402126	23	679
EPI_ISL_402127	33	767
EPI_ISL_402128	33	767
EPI_ISL_402129	55	890
EPI_ISL_402130	33	900
EPI_ISL_402131	44	678
EPI_ISL_402132	38	876
EPI_ISL_402133	56	890

Please paste a taxon table with the **sequence name**, **start and end position of gene**, separated by tab.

Run and Profess

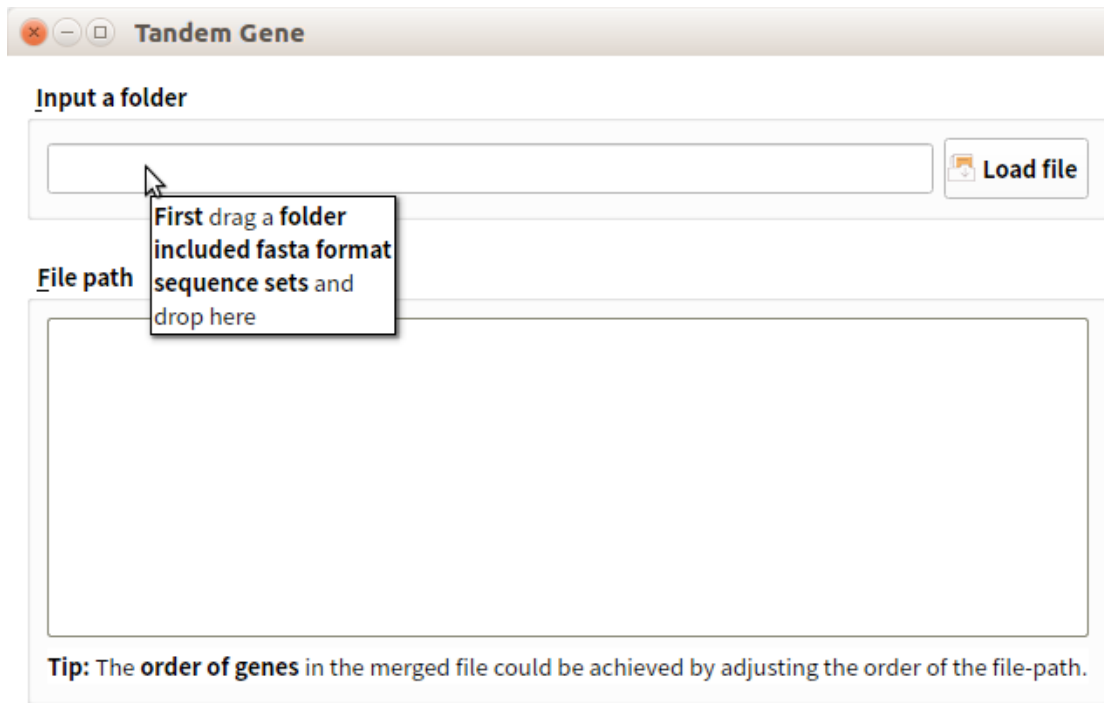
0%

Start

If users choose the options of ***"Equal range"***, BioAider will split all the sequences according to the same specified range.

3.1.5. *Combine Gene (Tandem Gene)*

This function is used to concatenate multiple gene sequences into one. Users can first put different genes dataset files into the same folder, and then drag the folder into the ***inputbox***, then click the ***"Load file button"*** import the file path of each genes datasets into ***textbox***. It should be pointed out that **the sequence names in different gene data sets should be consistent**, otherwise BioAider cannot be associated with them, but BioAider allows some data in a certain gene datasets to be missing and will represent them by gaps ("–").



Note, users can **modify the order of genes** in tandemly sequence by adjusting the sort of inputfile path in the **textbox**. **Of note**, all the sequences which are used for combined should be fasta format.

3.1.6. *Visual Gene Extractor*

This function is used to extract the sequences included specified gene from mixed coding gene sequence set , especially when these sequences datas are downloaded from NCBI database. Given that the same gene may have different manifestations in different studies, the textbox of **"Gene name"** could enter multiple names, and BioAider will extract the corresponding gene sequence which contain these gene names.

BioAider providers two versions, ***Visual Gene Extractor(Streamlit)*** and ***Visual Gene Extractor(Steady)***.The versions of ***Visual Gene Extractor(Streamlit)*** was developed based on ***Streamlit***, so if you use this version, you need to install Streamlit on your computer first. On the other hand, the ***Visual Gene Extractor(Steady)*** does not need any other environment.

Next, we demonstrate the use of ***Visual Gene Extractor(Steady)***.

Example 1 (The sequences data are directly downloaded from NCBI database, including some gene fragments of SARS-CoV-2):

Visual Gene Extractor(Steady)

Upload fasta file

Gene name(s)

gene=ORF6

Run

0% Click run

You can view the sequence name here

[location=join(227..13429,13429..21516)] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89516.1_2 [gene=ORF1ab] [protein=ORF1a polypeptide] [protein_id=QJR89516.1] [location=227..13444] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89517.1_3 [gene=S] [protein=surface glycoprotein] [protein_id=QJR89517.1] [location=21524..25345] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89518.1_4 [gene=ORF3a] [protein=ORF3a protein] [protein_id=QJR89518.1] [location=25354..26181] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89519.1_5 [gene=E] [protein=envelope protein] [protein_id=QJR89519.1] [location=26206..26433] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89520.1_6 [gene=M] [protein=membrane glycoprotein] [protein_id=QJR89520.1] [location=26484..27152] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89521.1_7 [gene=ORF6] [protein=ORF6 protein] [protein_id=QJR89521.1] [location=27163..27348] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89522.1_8 [gene=ORF7a] [protein=ORF7a protein] [protein_id=QJR89522.1] [location=27355..27720] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89523.1_9 [gene=ORF7b] [protein=ORF7b] [protein_id=QJR89523.1] [location=27717..27848] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89524.1_10 [gene=ORF8] [protein=ORF8 protein] [protein_id=QJR89524.1] [location=27855..28220] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89525.1_11 [gene=N] [protein=nucleocapsid phosphoprotein] [protein_id=QJR89525.1] [location=28235..29494] [gbkey=CDS]
 lc|MT451289.1_cds_QJR89526.1_12 [gene=ORF10] [protein=ORF10 protein] [transl_except=(pos:1..3,aa:Met)] [protein_id=QJR89526.1] [location=29519..29635] [gbkey=CDS]
 lc|MT451283.1_cds_QJR89443.1_1 [gene=ORF1ab] [protein=ORF1ab polypeptide] [exception=ribosomal slippage] [protein_id=QJR89443.1] [location=join(227..13429,13429..21516)] [gbkey=CDS]

Extract log

After uploading the sequence to BioAidrs as above, then we extract ORF6 gene sequence of SARS-CoV-2. Input a string **containing at least the gene name** to textbox of **"Gene name(s)"**, then click button of **"Run"**, run log as follows:

Visual Gene Extractor(Steady)

Upload fasta file

Gene name(s)

gene=ORF6

Run

100%

You can view the sequence name here

```

[location=join(227..13429,13429..21516)] [gbkey=CDS]
lc|MT451289.1_cds_QJR89516.1_2 [gene=ORF1ab] [protein=ORF1a polypeptide] [protein_id=QJR89516.1] [location=227..13444] [gbkey=CDS]
lc|MT451289.1_cds_QJR89517.1_3 [gene=S] [protein=surface glycoprotein] [protein_id=QJR89517.1] [location=21524..25345] [gbkey=CDS]
lc|MT451289.1_cds_QJR89518.1_4 [gene=ORF3a] [protein=ORF3a protein] [protein_id=QJR89518.1] [location=25354..26181] [gbkey=CDS]
lc|MT451289.1_cds_QJR89519.1_5 [gene=E] [protein=envelope protein] [protein_id=QJR89519.1] [location=26206..26433] [gbkey=CDS]
lc|MT451289.1_cds_QJR89520.1_6 [gene=M] [protein=membrane glycoprotein] [protein_id=QJR89520.1] [location=26484..27152] [gbkey=CDS]
lc|MT451289.1_cds_QJR89521.1_7 [gene=ORF6] [protein=ORF6 protein] [protein_id=QJR89521.1] [location=27163..27348] [gbkey=CDS]
lc|MT451289.1_cds_QJR89522.1_8 [gene=ORF7a] [protein=ORF7a protein] [protein_id=QJR89522.1] [location=27355..27720] [gbkey=CDS]
lc|MT451289.1_cds_QJR89523.1_9 [gene=ORF7b] [protein=ORF7b] [protein_id=QJR89523.1] [location=27717..27848] [gbkey=CDS]
lc|MT451289.1_cds_QJR89524.1_10 [gene=ORF8] [protein=ORF8 protein] [protein_id=QJR89524.1] [location=27855..28220] [gbkey=CDS]
lc|MT451289.1_cds_QJR89525.1_11 [gene=N] [protein=nucleocapsid phosphoprotein] [protein_id=QJR89525.1] [location=28235..29494] [gbkey=CDS]
lc|MT451289.1_cds_QJR89526.1_12 [gene=ORF10] [protein=ORF10 protein] [trans_except=(pos:1..3,aa:Met)] [protein_id=QJR89526.1] [location=29519..29635] [gbkey=CDS]
lc|MT451283.1_cds_QJR89443.1_1 [gene=ORF1ab] [protein=ORF1ab polypeptide] [exception=ribosomal slippage] [protein_id=QJR89443.1] [location=join(227..13429,13429..21516)] [gbkey=CDS]
lc|MT451283.1_cds_QJR89444.1_2 [gene=ORF1ab] [protein=ORF1a polypeptide] [protein_id=QJR89444.1] [location=227..13444] [gbkey=CDS]

```

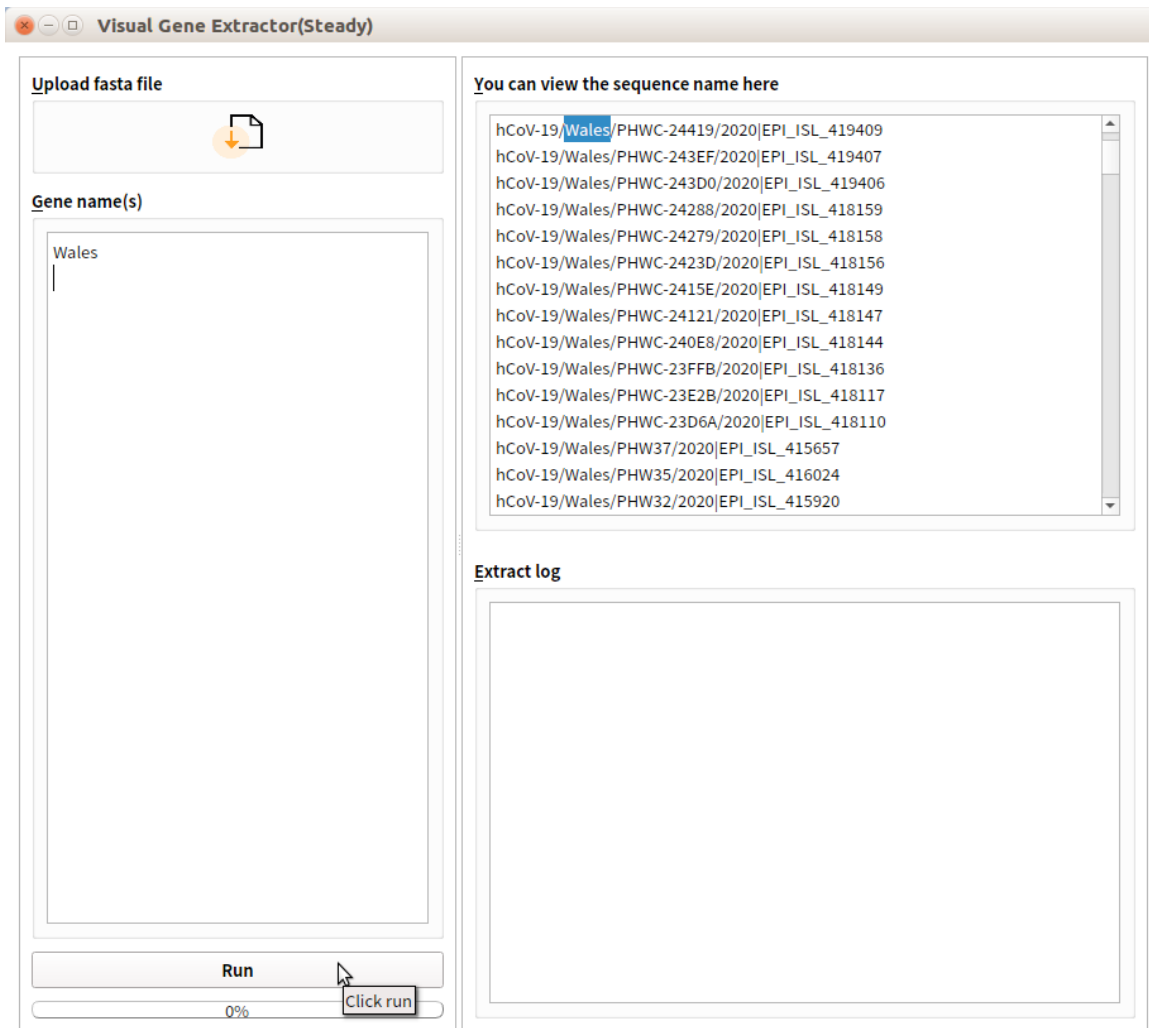
Extract log

GenID list	Have been extracted gene
MT451283.1	[1]QJR89449.1_7_gene=ORF6
MT450923.1	[1]QJR85129.1_7_gene=ORF6
MT419838.1	[1]QJ54150.1_7_gene=ORF6
MT419850.1	[1]QJR88513.1_7_gene=ORF6
MT451205.1	[1]QJR89329.1_7_gene=ORF6
MT451273.1	[1]QJ54315.1_7_gene=ORF6
MT419852.1	[1]QJ54019.1_7_gene=ORF6
MT419827.1	[1]QJQ27858.1_7_gene=ORF6
MT434758.1	[1]QJR95629.1_7_gene=ORF6
MT451798.1	[1]QJ54280.1_7_gene=ORF6
MT419849.1	[1]QJ54209.1_7_gene=ORF6
MT419843.1	[1]QJ54174.1_7_gene=ORF6
MT419840.1	[1]QJR89521.1_7_gene=ORF6
MT451289.1	[1]QJR89521.1_7_gene=ORF6

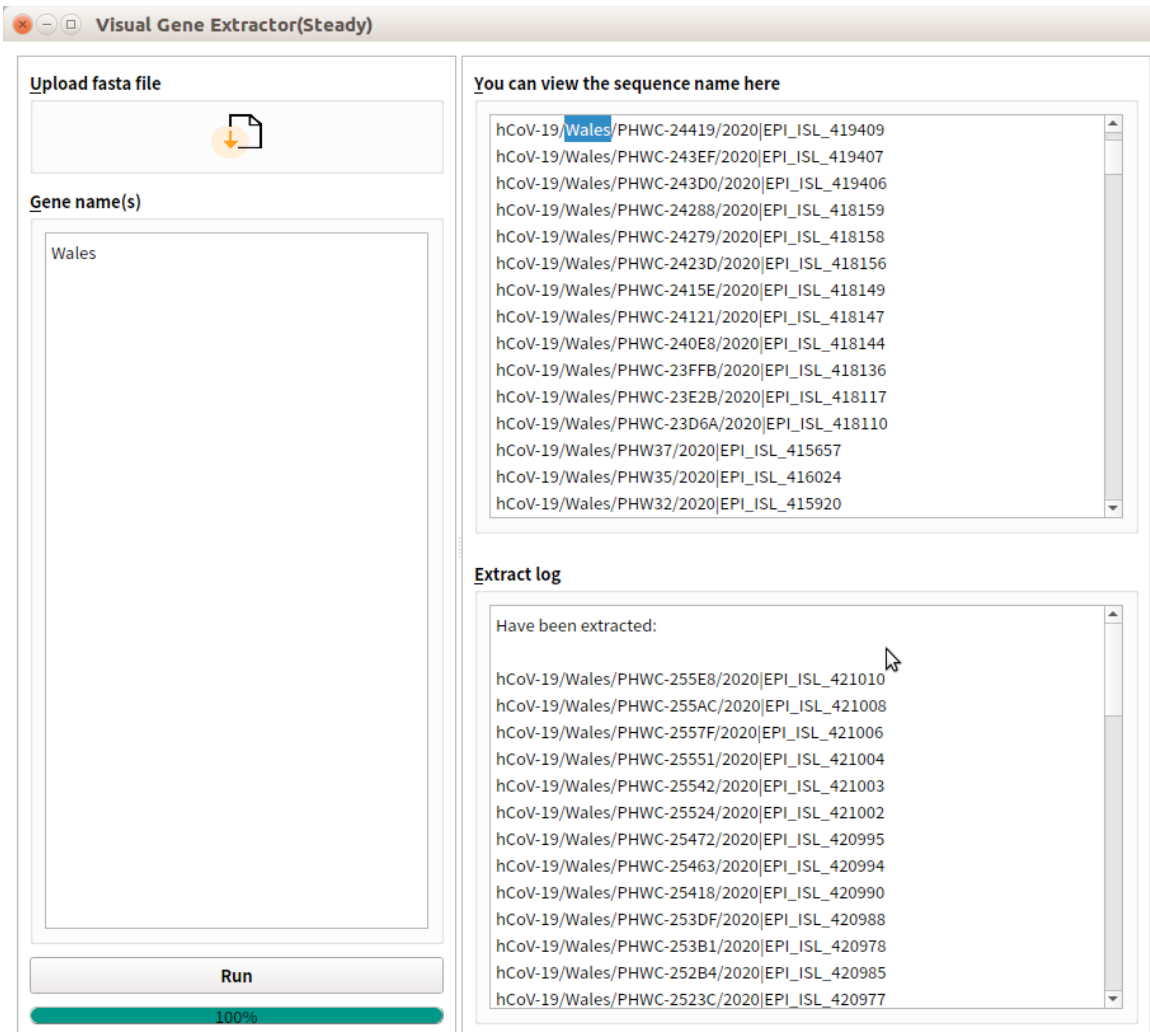
Extracted counts:13

If some gene fragments are not extracted (as shown by the arrow), the possible reason is that the strain does not contain this gene fragment or the gene owns other names in some sequences. If it is the second case, you can append other names of this gene to the next line of **"Gene name(s)"** textbox.

Example 2 (Arbitrary fasta sequence):



As shown above, if you want to extract these sequences which containing the tags of "*Wales*", BioAider could accomplish it well.



The extracted sequence will be saved in the directory where the input file is located.

3.1.7. Fast Annotation

For these strain sequences from the same or highly related species, their nucleotide identity is usually relatively higher. Therefore, the sequences annotation could be based on the gene information of the reference sequence after multi-sequence alignment.

Fast Annotation

Open file

+

Tip: The **reference sequence** for annotation have to be placed **first in the sequence set**

Information of gene

#Example
ORF1ab,AAAAGGGGCCCTT,AAAAACCTTTCCC
S,TTTAAAAGGCCTTC,GGGGTATTTATATATCTTAA
ORF3a,AAGCCCTCTGCTCGTCG,TTGTGAGTGTGGA
|

Please pasta a table with the
names, start and end
character string of gene of
first seq in data set,
separated by ", ".

BioAider provides a quickly sequence annotation function, users can import the aligned complete genome sequence set (fasta format file), and adjust the reference sequence for annotation to the forefront of the file. Paste the gene information of reference sequence in aligned sets, name, starting string and end string into the textbox, **separated by ", "**. Then batch abstract genes.

Note that the start string or end string of the gene is not limited in length, but it is required to be unique in the reference sequence. Besides, the higher of similarity among sequences, the higher accuracy of the annotation.

3.2. Similar Analysis

3.2.1. Sequence Identity Matrix

By inputting the aligned sequence datasets in fasta format, and a pairwise sequence identity matrix can be generated. This function contains two different

modes: nucleotide or amino acid sequence identity matrix ("*Single nt or aa*"), nucleotide plus amino acid sequence identity matrix ("*Combination nt and aa*").

Sequence Identity Matrix

Open file

/home/gong/BioAider_develop/src/dist/Example/Test-aligned.fasta

Type of matrix

☐ Single nt or aa ☒ Combination nt and aa (nt/aa)

Condense gap?

☐ Yes

Select a code table

Codon Table: 1.The Stand

Run and Profess

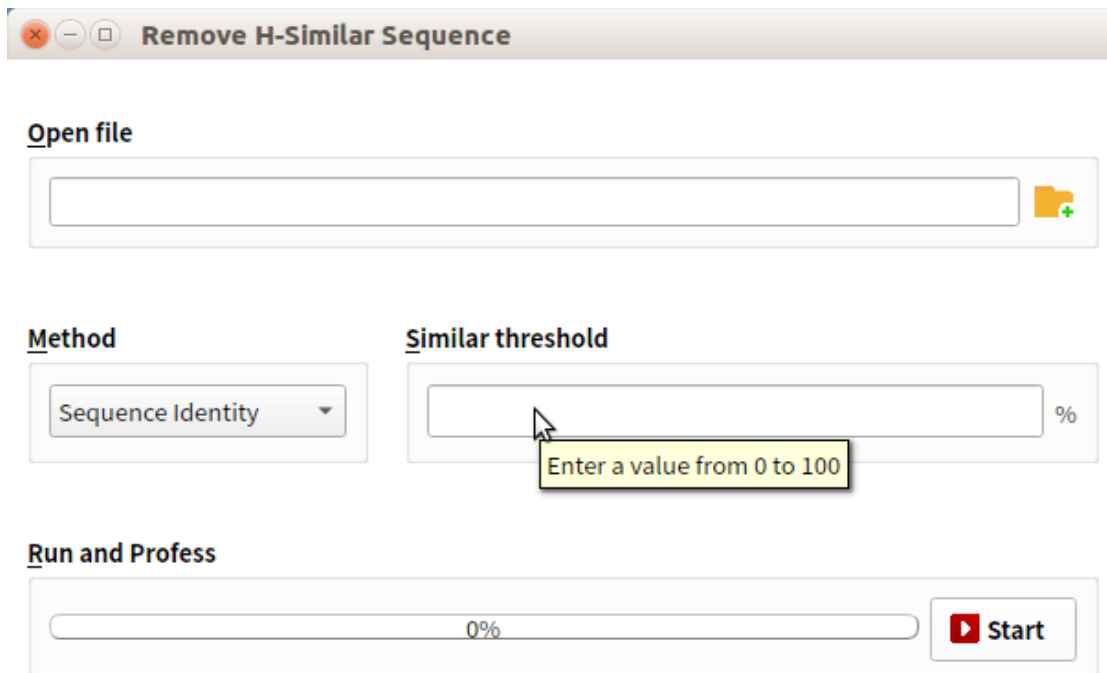
0%

Start

It should be noted that if the "*Combination nt and aa*" is selected, the inputted sequences should be aligned based on codon method in advance. In order to better fit the variation characteristics, BioAider provides the "*Condense gap*" function. If the option was selected, the program will treat every three consecutive inserted or deleted bases as one.

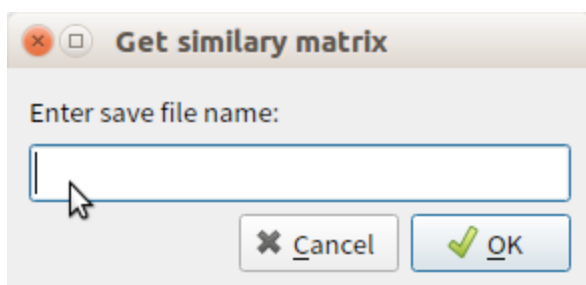
3.2.2 Remove H-Similar Sequence

This function could remove highly similar sequences and keep one by specifying the threshold of similarity ("*Similar threshold*"). BioAider provides 6 different methods for calculating the similarity of sequences.



It should be noted that the *"Sequence Identity"* and *"Hamming"* methods require the input sequences data are aligned, and we suggest that the sequences datasets for remaining 4 methods better not be pre-aligned, because these algorithm own alignment function. If *"Similar threshold"* is seted to 100, the function of eliminateing duplicate sequences will be turned on. **Note**, if the *"Similar threshold"* is set to 100, no matter what algorithm is selected, it is the same because the program adopts another efficient processing mechanism.

If you want to obtain the sequence similarity matrix calculated by the above 6 methods, you can click the right button of mouse in any region of the program interface to call up the functional menu.

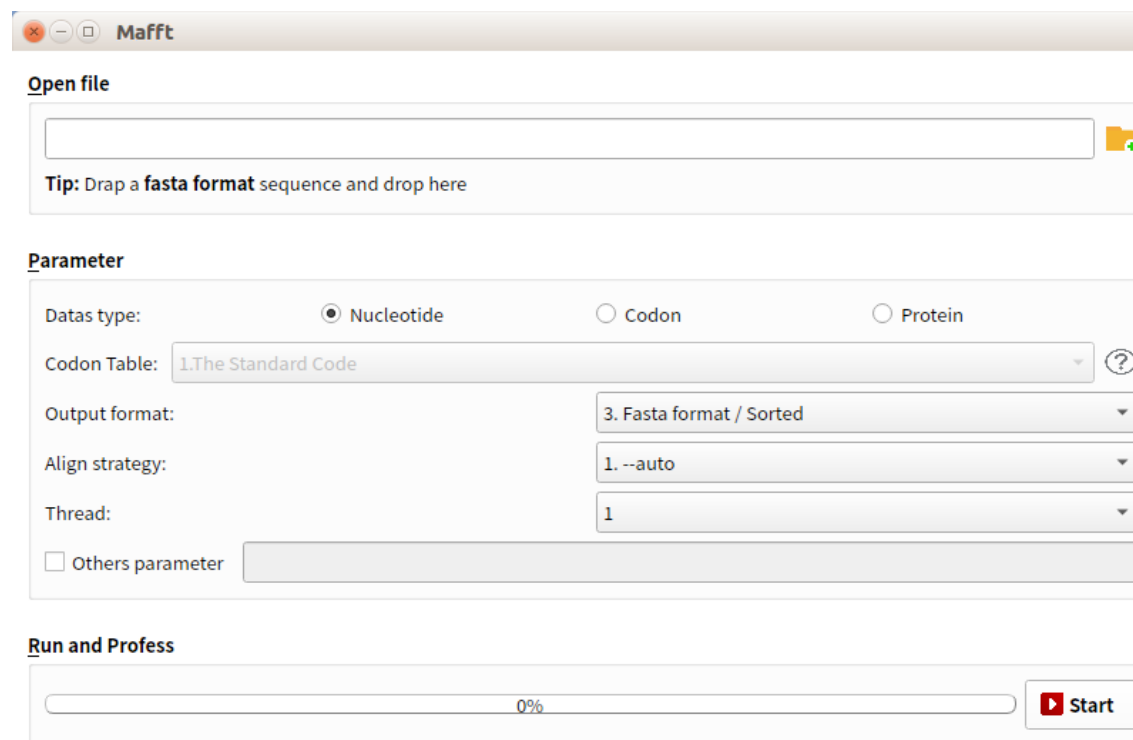


3.3. Align tools

Multiple–Sequence–Alignment (MSA) is the most common analysis in sequence processing, most classic MSA software runs as a command symbol. It is very inconvenient for non–bioinformatics analysts. **BioAider packed three MSA software** (Mafft, Muscle and Clustal–Omega) in the graphical interface, **and provided translation–alignment additionally based on multiple sets of codon tables.**

3.3.1. *Mafft*

Mafft is a very popular MSA software with higher comparison accuracy, and its comparison speed is also relatively good. Some common parameter are encapsulated into the graphical interface in BioAider, and other parameters also could be add flexibly. More detailed information about Mafft could be got from <https://mafft.cbrc.jp/alignment/software/>.



The screenshot shows the Mafft interface with the following sections:

- Open file:** A text input field with a folder icon and a tip: "Tip: Drag a **fasta format** sequence and drop here".
- Parameter:**
 - Datas type:** Radio buttons for Nucleotide (selected), Codon, and Protein.
 - Codon Table:** A dropdown menu showing "1.The Standard Code" with a help icon.
 - Output format:** A dropdown menu showing "3. Fasta format / Sorted".
 - Align strategy:** A dropdown menu showing "1. --auto".
 - Thread:** A dropdown menu showing "1".
 - Others parameter:** A checkbox labeled "Others parameter" followed by a text input field.
- Run and Profess:** A progress bar showing "0%" and a "Start" button with a play icon.

3.3.2. *Muscle*

The comparison rate and accuracy of Muscle are good, according to the instruction manual of Muscle, setting ***"Maxiters"*** to 1 or 2 will significantly speed up the operation.

More detailed please reference

<http://petrov.stanford.edu/software/src/muscle3.6/muscle3.6.html>.

Open file

Tip: Drap a sequence sets and drop here

Parameter

Datas type: ☒ Nucleotide ☐ Codon ☐ Protein

Codon Table: 1.The Standard Code

Output format: 1. Fasta format

Maxiters: 16

☐ Others parameter

Run and Profess

0% **Start**

3.3.3. *Clustal–Omeg*

As a relatively classic MSA software, *Clustal* has a broad user base. As the latest addition to the Clustal family. Clustal–Omega offers a significant increase in scalability over previous versions, more detailed reference <http://www.clustal.org/omega/>.

Open file

Tip: Drap a sequence sets and drop here

Parameter

Datas type: ☒ Nucleotide ☐ Codon ☐ Protein

Codon Table: 1.The Standard Code

Output format: 1. Fasta format

Thread:

☐ Output guidetree? ☐ Others parameter

Run and Profess

0%

3.4. Mutation Tools

3.4.1. Mutation Analysis

This function could be used for analysis of the **mutations characteristicson on large numbers of sequenced strains**. The sequence datas for analysis needs to be aligned in advance, and they could be nucleotides, proteins (amino acid) sequences or simply coding gene fragments. For nucleotides and proteins sequences, BioAider could summarizes all the mutation sites with corresponding frequency and strains.

Of course, if the datas is codon gene, BioAider provides multiple sets of different codon tables for users, and could scan each condon sites in aligned sequence datasets, and identifies the type of mutation, including synonymous, non-synonymous, insertions and deletions and early termination. Finally, BioAider will automatically summarize and output the relevant analysis results.

Note: The codon gene sequences for mutations analysis have to be aligned by translation-alignment methon in advance, It is worth mentioning that BioAider packed three multiple-sequence-alignment software (mafft, muscle and clsutal-omega) in the graphical interface, and provided translation-alignment additionally.

Whether it's nucleotides or amino acids or coding genes, BioAider could plot the frequency distribution graph for mutation sites through specifing groups of substitution frequency in custom.

Eaxmple of mutations analysis for aligned SARS-CoV-2 ORF3a gene sequences. First, create frequency grouping in a table editor:

Liberation Sans			
小二			
B I U			
T			
B1:C6			
fx Σ = 3040			
	A	B	C
1	Group1	0	1
2	Group2	1	5
3	Group3	5	10
4	Group4	10	15
5	Group5	15	20
6	Group6	20	3040
7			
8			

The each groups of substitution frequency contains start value and end value which are separated by tab symbol. **Note, the start value** of each group is not included in the range of frequency, **and the frequencies of different groups need to be consecutive integers.**

Then copy them to the textedit box of BioAider, and select "Codon" single button in "Datas type":

Mutation Analysis

Open file

/home/gong/BioAider_develop/src/dist/Example/SARS-CoV-2_ORF3a_aligned.fas

Tip: The first sequence in the data set will be used as the reference sequence for mutation analysis

Parameter

Datas type:

☐ Nucleotide

☒ Codon

☐ Protein

Codon Table:

1.The Standard Code

Output substitution frequency distribution?

☒ Yes

Groups of substitution frequency

01

15

510

1015

1520

203040

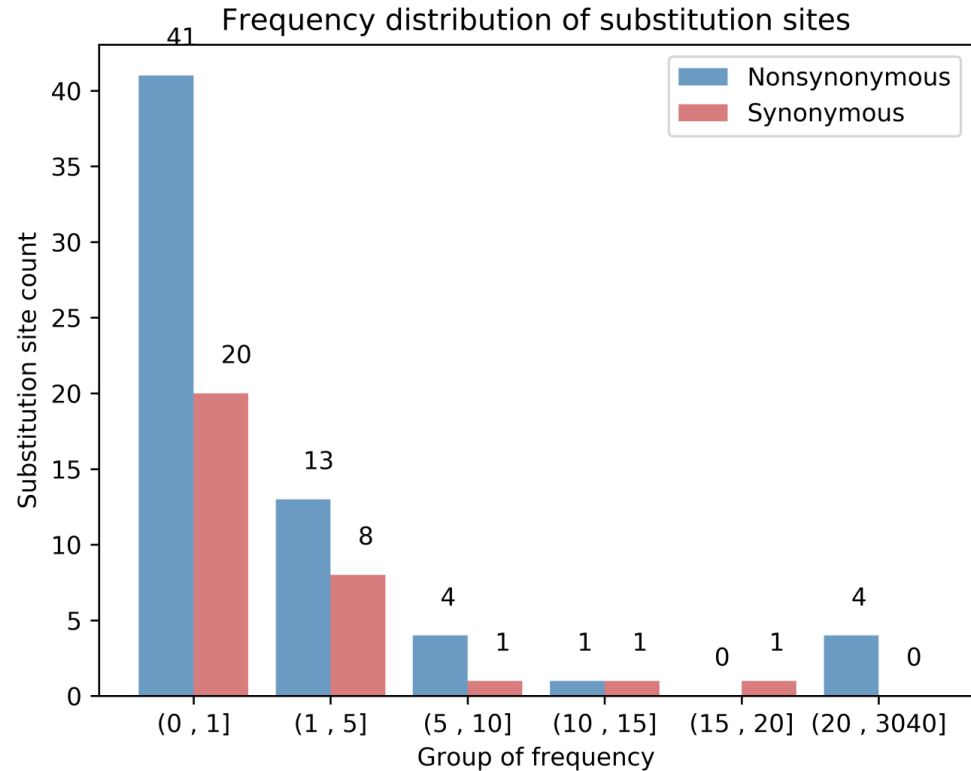
Please pasta the frequency groups.

Tip: Only sys and non-sys substitution are considered for calculation.

After the run is over, these analysis result could be found in the directory where the source file is located, you could scan the **_mutation site summary* file then know the overall variation and mutation hotspots.

Reference sequence: hCoV-19/Wuhan/VDC-HB-01/2019/EPI_ISL_402119						
Site (codon)	Index of nt	mutation type	mutation	changed in properties of aa?	changed type of properties	sequence count
8	1	Nonsynonymous	TTC to CTC (F to L)	No	No	1
8	3	Synonymous	TTC to TTT	No	No	1
13	1	Nonsynonymous	GTA to TTA (V to L)	No	No	8
14	2	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	7
15	3	Nonsynonymous	TTG to TTT (L to F)	No	No	1
19	3	Synonymous	GAA to GAG	No	No	1
26	1	Nonsynonymous	TCA to CCA (S to P)	Yes	(polar,none-charge) to (non-polar,none-charge)	1
27	3	Synonymous	GAT to GAC	No	No	3
28	3	Synonymous	TTT to TTC	No	No	1
31	1	Nonsynonymous	GCT to ACT (A to T)	Yes	(non-polar,none-charge) to (polar,none-charge)	2
34	2	Nonsynonymous	ACG to ATG (T to M)	Yes	(polar,none-charge) to (non-polar,none-charge)	1
34	3	Synonymous	ACG to ACT	No	No	1
36	3	Synonymous	CCG to CCT	No	No	15
36	3	Synonymous	CCG to CCA	No	No	2
43	3	Synonymous	TTT to TTT	No	No	1
46	1	Nonsynonymous	CTT to TTT (L to F)	No	No	2
49	1	Nonsynonymous	GGC to TGC (G to C)	No	No	1
49	3	Synonymous	GGC to GGT	No	No	2
50	2	Nonsynonymous	GTT to GCT (V to A)	No	No	1
54	1	Nonsynonymous	GCT to TCT (A to S)	Yes	(non-polar,none-charge) to (polar,none-charge)	1
54	3	Synonymous	GCT to GCG	No	No	1
54	2	Nonsynonymous	GCT to GTT (A to V)	No	No	3
56	2	Nonsynonymous	TTT to TGT (F to C)	Yes	(non-polar,none-charge) to (polar,none-charge)	1
57	3	Nonsynonymous	CAG to CAT (Q to H)	Yes	(polar,none-charge) to (polar,positive-charge)	578
59	2	Nonsynonymous	GCT to GAT (A to D)	Yes	(non-polar,none-charge) to (polar,negative-charge)	2
60	3	Synonymous	TCC to TCT	No	No	4
60	2	Nonsynonymous	TCC to TTC (S to F)	Yes	(polar,none-charge) to (non-polar,none-charge)	1
61	3	Nonsynonymous	AAA to AAC (K to N)	Yes	(polar,positive-charge) to (polar,none-charge)	2
65	3	Synonymous	CTC to CTT	No	No	1
67	3	Nonsynonymous	AAG to AAT (K to N)	Yes	(polar,positive-charge) to (polar,none-charge)	5
67	3	Nonsynonymous	AAG to AAC (K to N)	Yes	(polar,positive-charge) to (polar,none-charge)	2
68	1	Synonymous	AGA to CGA	No	No	1
74	1	Nonsynonymous	TCC to GCC (S to A)	Yes	(polar,none-charge) to (non-polar,none-charge)	1
75	1	Nonsynonymous	AAG to GAG (K to E)	Yes	(polar,positive-charge) to (polar,negative-charge)	2
76	1	Nonsynonymous	GGT to AGT (G to S)	No	No	1
78	1	Nonsynonymous	CAC to TAC (H to Y)	Yes	(polar,positive-charge) to (polar,none-charge)	1
84	3	Synonymous	CTG to CTT	No	No	1
85	1	Synonymous	TTG to CTG	No	No	1
87	2	Nonsynonymous	TTT to TAT (F to Y)	Yes	(non-polar,none-charge) to (polar,none-charge)	1
88	2	Nonsynonymous	GTA to GCA (V to A)	No	No	2

You could also know the number of mutation sites under each mutation frequency group through view **_substitution frequency distribution.png*.



It is not difficult to find that more than half of the mutation sites only appear in a single strain, although there are many mutation sites in ORF3a gene. Of course, BioAider additionally provides vector graphics (**_substitution frequency distribution.pdf*), users can edit them and facilitate publication.

Besides, users could obtain the corresponding mutant strains of these variant sites in the detailed **_log.txt* file.

seq name							
A	B	C	D	E	F	G	H
1	Reference sequence: hCoV-19/Wuhan/VDC-HB-01/2019/EPI_ISL_402119						
2							
Site (codon)	Index of nt	seq name	mutation type	mutation	changed in properties of aa?	changed type of properties	
4	8	11337hCoV-19/Netherlands/NA_30/2020/EPI_ISL_415487	Nonsynonymous	TTC to CTC (F to L)	No	No	
5	8	11985hCoV-19/France/Poligny_1733/2020/EPI_ISL_416745	Synonymous	TTC to TTT			
6	13	14749hCoV-19/USA/WA-5119/2020/EPI_ISL_417172	Unknown	GTA to NTA			
7	13	1562hCoV-19/USA/UT-00023/2020/EPI_ISL_418965	Nonsynonymous	GTA to TTA (V to L)	No	No	
8	13	11799hCoV-19/Iceland/127/2020/EPI_ISL_417727	Nonsynonymous	GTA to TTA (V to L)	No	No	
9	13	12241hCoV-19/England/SHEF-C044C/2020/EPI_ISL_420193	Nonsynonymous	GTA to TTA (V to L)	No	No	
10	13	12298hCoV-19/England/20142061804/2020/EPI_ISL_420772	Nonsynonymous	GTA to TTA (V to L)	No	No	
11	13	12331hCoV-19/England/20126007102/2020/EPI_ISL_418751	Nonsynonymous	GTA to TTA (V to L)	No	No	
12	13	12332hCoV-19/England/20126006802/2020/EPI_ISL_418749	Nonsynonymous	GTA to TTA (V to L)	No	No	
13	13	12978hCoV-19/Australia/VIC303/2020/EPI_ISL_419994	Nonsynonymous	GTA to TTA (V to L)	No	No	
14	13	13006hCoV-19/Australia/VIC275/2020/EPI_ISL_419968	Nonsynonymous	GTA to TTA (V to L)	No	No	
15	14	2089hCoV-19/France/CVL2000/2020/EPI_ISL_418222	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	
16	14	2093hCoV-19/France/Bourg-en-Bresse_06678/2020/EPI_ISL_416757	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	
17	14	2450hCoV-19/Canada/431/2020/EPI_ISL_420847	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	
18	14	2545hCoV-19/Canada/NB_6/2020/EPI_ISL_418811	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	
19	14	2641hCoV-19/Belgium/ULG-9739/2020/EPI_ISL_418663	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	
20	14	2649hCoV-19/Belgium/ULG-9647/2020/EPI_ISL_418653	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	
21	14	2811hCoV-19/Belgium/FAE-030948/2020/EPI_ISL_420445	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	
22	15	2716hCoV-19/Belgium/SR-0319112/2020/EPI_ISL_420369	Nonsynonymous	TTG to TTT (L to F)	No	No	
23	19	1535hCoV-19/Italy/TE6222/2020/EPI_ISL_420583	Unknown	GAA to GAW			
24	19	2956hCoV-19/Australia/VIC324/2020/EPI_ISL_420015	Synonymous	GAA to GAG			
25	19	3142hCoV-19/Australia/VIC135/2020/EPI_ISL_419828	Synonymous	GAA to GAG			
26	20	2252hCoV-19/England/SHEF-C036D/2020/EPI_ISL_420206	Unknown	ATC to ATN			
27	26	2552hCoV-19/Canada/BC_8896915/2020/EPI_ISL_418854	Nonsynonymous	TCA to CCA (S to P)	Yes	(polar,none-charge) to (non-polar,none-charge)	
28	27	1381hCoV-19/Malaysia/188407/2020/EPI_ISL_417918	Synonymous	GAT to GAC			
29	27	3102hCoV-19/Australia/VIC178/2020/EPI_ISL_419875	Synonymous	GAT to GAC			
30	27	3133hCoV-19/Australia/VIC144/2020/EPI_ISL_419841	Synonymous	GAT to GAC			
31	28	1178hCoV-19/USA/WI-13/2020/EPI_ISL_417516	Synonymous	TTT to TTC			
32	31	1818hCoV-19/Hungary/mb1/2020/EPI_ISL_416426	Nonsynonymous	GCT to ACT (A to T)	Yes	(non-polar,none-charge) to (polar,none-charge)	
33	31	13201hCoV-19/Australia/NSW47/2020/EPI_ISL_417402	Nonsynonymous	GCT to ACT (A to T)	Yes	(non-polar,none-charge) to (polar,none-charge)	
34	34	76hCoV-19/Wales/PHWC-2415E/2020/EPI_ISL_418149	Nonsynonymous	ACG to ATG (T to M)	Yes	(polar,none-charge) to (non-polar,none-charge)	
35	34	300hCoV-19/USA/WA-UW215/2020/EPI_ISL_417364	Synonymous	ACG to ACT			
36	36	116hCoV-19/USA/WI-77/2020/EPI_ISL_421335	Synonymous	CCG to CCT			
37	36	120hCoV-19/USA/WI-73/2020/EPI_ISL_421331	Synonymous	CCG to CCT			
38	36	122hCoV-19/USA/WI-71/2020/EPI_ISL_421329	Synonymous	CCG to CCT			
39	36	123hCoV-19/USA/WI-70/2020/EPI_ISL_421328	Synonymous	CCG to CCT			
40	36	124hCoV-19/USA/WI-69/2020/EPI_ISL_421327	Synonymous	CCG to CCT			
41	36	126hCoV-19/USA/WI-67/2020/EPI_ISL_421325	Synonymous	CCG to CCT			
42	36	127hCoV-19/USA/WI-66/2020/EPI_ISL_421324	Synonymous	CCG to CCT			
43	36	132hCoV-19/USA/WI-61/2020/EPI_ISL_421319	Synonymous	CCG to CCT			

Of note, if these sequences are much divergent, such as from different family enver order and contain a lot of gaps ("-") in the aligned sequence, I usually don't recommend using them for mutation analysis. On the one hand, they would make a lot of calculations, on the other hand, they are inherently highly variable and have no value of analysis.

But if you still want to study their variation, it is recommended to use the following function of *"Site Counter"*.

3.4.2. Site Counter

This function could summary the type, count and proportion of bases (or amino acids) at each site for the aligned sequence datasets. In addition, BioAider will output a consensus sequence based on the highest proportion base (or amino acid) in each site.

The screenshot shows the 'Site Counter' web application. It has a title bar with standard window controls. Below the title bar, there are three main sections: 'Open file', 'Data type', and 'Profess'. The 'Open file' section contains a text input field with the path '/home/gong/BioAider_develop/src/dist/Example/Test-aligned.fasta' and a folder icon. The 'Data type' section has two radio buttons: 'Nucleotide' (selected) and 'Protein'. The 'Profess' section features a progress bar at 0% and a 'Start' button with a red play icon. A tooltip 'Click run' is visible over the 'Start' button.

For DNA sequence datasets, the one of results (**_site_count.csv*) was as follows:

I12		f _x	Σ	=				
	A	B	C	D	E	F	G	H
1		site1	site2	site3	site4	site5	site6	site7
2	A	29	0	0	0	0	0	0
3	G	0	0	29	0	0	0	29
4	C	0	0	0	0	0	0	0
5	T	0	29	0	29	29	29	0
6	-	0	0	0	0	0	0	0
7								
8								

3.4.3. Site Scree

This function is used to extract the sequences with corresponding base (or amino acid) in *specified one or more* site(s). It is very useful for studying whether there is linkage inheritance among different gene sites.

Site Scree

Open file

/home/gong/BioAider_develop/Example/test-scan-seq.fas

Information of site(s)

2,T
10,T
|

Please pasta **position and nt(or aa)**, separated by ",". For example: 174,A

4. Test Datas

Examples and test are available

at <https://github.com/ZhijianZhou01/BioAider/tree/master/Example>.