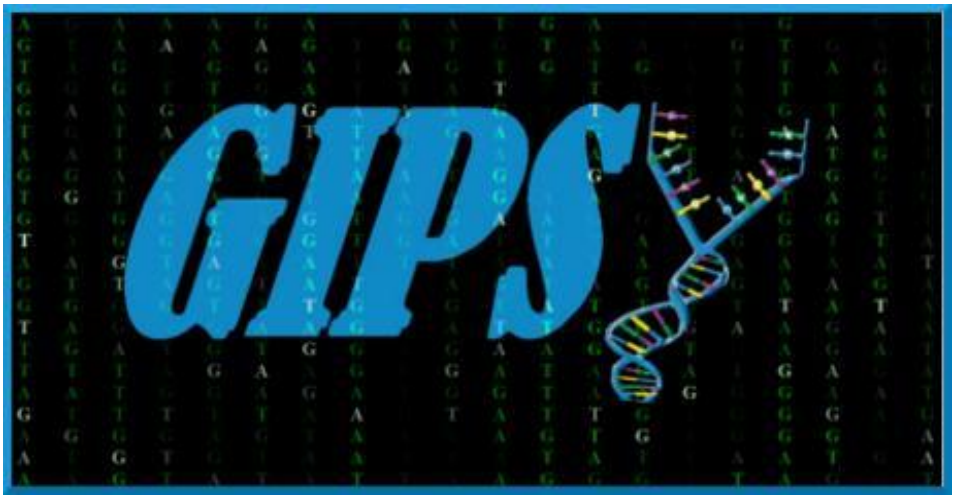


GIPSy: Genomic Island Prediction Software

Version 1.1.2

Summary

I. What is GIPSy?

GIPSy is a software for accurate prediction of genomic islands into the classes: pathogenicity islands (PAIs), resistance Islands (RIs), metabolic Islands (MIs) and symbiotic Islands (SIs).

II. Importance of genomic islands

Bacteria are highly diverse organisms that are able to adapt to a broad range of environments and hosts due to their high genomic plasticity. Horizontal gene transfer plays a pivotal role in this genome plasticity and evolution by leaps through the incorporation of large blocks of genome sequences, ordinarily known as genomic islands (GEIs). GEIs may harbor genes encoding virulence, metabolism, antibiotic resistance and symbiosis-related functions, namely pathogenicity islands (PAIs), metabolic islands (MIs), resistance islands (RIs) and symbiotic islands (SIs).

III. Method summary

GIPSy predicts GEIs based on commonly shared features: genomic signature deviation (G+C content and codon usage); presence of tranposase genes; virulence, metabolism, antibiotic resistance, or symbiosis factors; flanking tRNA genes; and absence in other organisms of the same genus or closely related species. Eight steps are necessary to evaluate the presence of these genomic features in GIPSy.

IV. GIPSy input formats

GIPSy accepts complete genomes only of bacteria in embl or genbank formats.

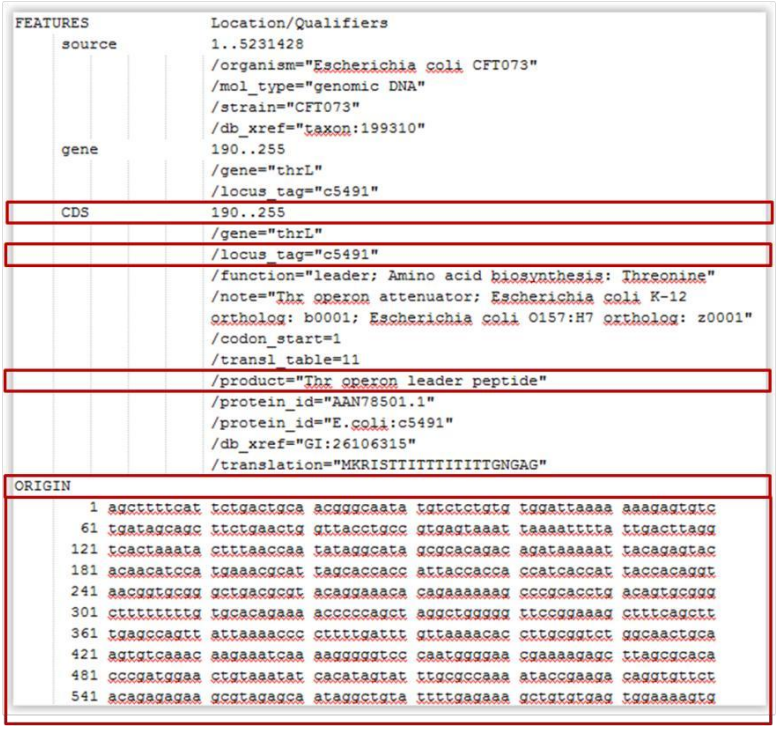
IV.1 EMBL format (.embl)

The embl format has the following structure, where the regions highlighted with red boxes are important for GIPSy analyses.

```
FT   source          1..3309401
FT                   /organism="Corynebacterium glutamicum ATCC 13032"
FT                   /strain="ATCC 13032"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:196627"
FT   CDS             1..1575
FT                   /codon_start=1
FT                   /transl_table=11
FT                   /locus_tag="Cg10001"
FT                   /product="ATPase involved in DNA replication initiation"
FT                   /translation="MSQNSSSLLETWRQVVADLTTLSQQADSGFDPLTPTQRAYLNLTK
FT                   PIAIVDGYAVLSTPNAMAKNVIENDLGDALTRVLSLRMGRSFSLAVSVEPEQEIPETPA
FT                   QQEFKYQPDAPVISSNKAPKQYEVGGRGEASTSDGWERTHSAPAPEPHPAPIADPEPEL
FT                   ATPQRIPRETPAHNPNREVSLNPKYTFESFVIGPFNRFANAAAVAVAESPAKAFNPLFI
FT                   SGGSGLGKTHLLHAVGNYAQELQPGLRIKYVSSEEFTNDYINSVRDDRQETFKRRYRNL
FT                   DILMVDDIQFLAGKEGTQEEFFHTFNALHQADKQIILSSDRPPKQLTTLEDRLRTRFEG
FT                   GLITDIQPPDLETRIAILMKKAQTDGTHVDREVLELIASRFESSIRELEGALIRVSAYS
FT                   SLINQPIDKEMAIVALRDILPEPEDMEITAPVIMEVTAEYFEISVDTLRGAGKTRAVAH
FT                   ARQLAMYLCRELTDMSLPKIGDVFGGKDHTTVMYADRKIRQEMTEKRDTYDEIQQLTQL
FT                   IKSRGRN"
XX
SQ   Sequence 3309401 BP; 764350 A; 894542 C; 886255 G; 764254 T; 0 other;
     gtgagccaga actcatcttc tttgctcgaa acctggcgcc aagttgttgc cgatctcaca        60
     actttgagcc agcaagcgga cagtggattc gacccattga cgccaactca acgtgcatat       120
     ttgaacctga cgaagccgat tgccatcgtc gatggctacg ccgtgctgtc cacacccaac       180
     gcgatggcaa aaaatgtcat tgaaaacgat ttgggcgatg ctttgacccg tgtgttgtcg       240
     ctgcgcatgg gccgatcatt cagcttggct gtcagtgtgg agcctgagca ggaaattcca       300
     gaaacccgag ctcagcagga gtttaaatat cagcctgacg cacctgtgat ctctttccaac      360
     aaggcgccaa agcagtatga agttggtggt cggggagagg cgtcgacaag cgacggctgg       420
     gaacgtaccc actctgcacc ggctcccgag ccgcacccgg cacctatcgc cgatcctgag       480
     ccagagctgg ccacccccgca gcgcattccg cgcgaaaccc cagctcacaa ccctaatcgg      540
     gaagtgtccc tcaacccgaa atacactttt gaaagcttcg tgatcgggcc gttcaaccgt       600
     ttcgccaatg cagccgcagt tgctgtggcg gaaagcccag cgaaagcttt caacccgctg       660
```

IV.2 Genbank format (.genbank, .gb, .gbk)

Although presenting similar information, the genbank file presents a different structure, as shown below. Likewise the embl format example, the important features are highlighted by red boxes in the figure.



```
FEATURES             Location/Qualifiers
     source          1..5231428
                     /organism="Escherichia coli CFT073"
                     /mol_type="genomic DNA"
                     /strain="CFT073"
                     /db_xref="taxon:199310"
     gene            190..255
                     /gene="thrL"
                     /locus_tag="c5491"
     CDS             190..255
                     /gene="thrL"
                     /locus_tag="c5491"
                     /function="leader; Amino acid biosynthesis: Threonine"
                     /note="Thr operon attenuator; Escherichia coli K-12
                     ortholog: b0001; Escherichia coli O157:H7 ortholog: z0001"
                     /codon_start=1
                     /transl_table=11
                     /product="Thr operon leader peptide"
                     /protein_id="AAN78501.1"
                     /protein_id="E.coli:c5491"
                     /db_xref="GI:26106315"
                     /translation="MKRISTTIITTTITITTGNGAG"
ORIGIN
        1 agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
       61 tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg
      121 tcactaaata ctttaaccaa tataggcata gcgcacagac agataaaaat tacagagtac
      181 acaacatcca tgaaacgcat tagcaccacc attaccacca ccatcaccat taccacaggt
      241 aacggtgcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg
      301 cttttttttg tgcacagaaa acccccagct aggctggggg ttccggaaag ctttcagctt
      361 tgagccagtt attaaaaccc cttttgattt gttaaaacac cttgcggtct ggcaactgca
      421 agtgtcaaac aagaaatcaa aaggggtcc caatggggaa cgaaaagagc ttagcgcaca
      481 cccgatggaa ctgtaaatat cacatagtat ttgcgccaaa ataccgaaga caggtgttct
      541 acagagagaa gcgtagagca ataggctgta ttttgagaaa gctgtgtgag tggaaaagtg
```

1. Installation process:

GIPSy may be used both in Windows- and Linux-based platforms. It only requires a Java Virtual Machine 1.7.0_51-b13 or higher to work. However, **we strongly advise the use of openjdk** instead of the Oracle version of java virtual machine **when working in linux-based machines** as the Oracle version may result in some exceptions during the analyses. The download of all databases and dependencies as well as the environment setup are performed by GIPSy.Installer.jar.

GIPSy.Installer.jar may be run in Windows by simply double clicking, whereas in Linux it only requires a simple command line, as follow:

java –jar GIPSy.Installer.jar

a) choose the appropriate folder for installation:

b) click install:

c) the installer will automatically download the databases, dependencies, GIPSy.jar and set up all the environment:

d) one shortcut will be created on Desktop.

OBS.: If it is not created, a prompt will appear warning that the Desktop folder was not found. In this case, simply go to the installation folder defined on the beginning of the installation process and execute GIPSy.jar from within this folder or create a Desktop shortcut.

2. Running GIPSy

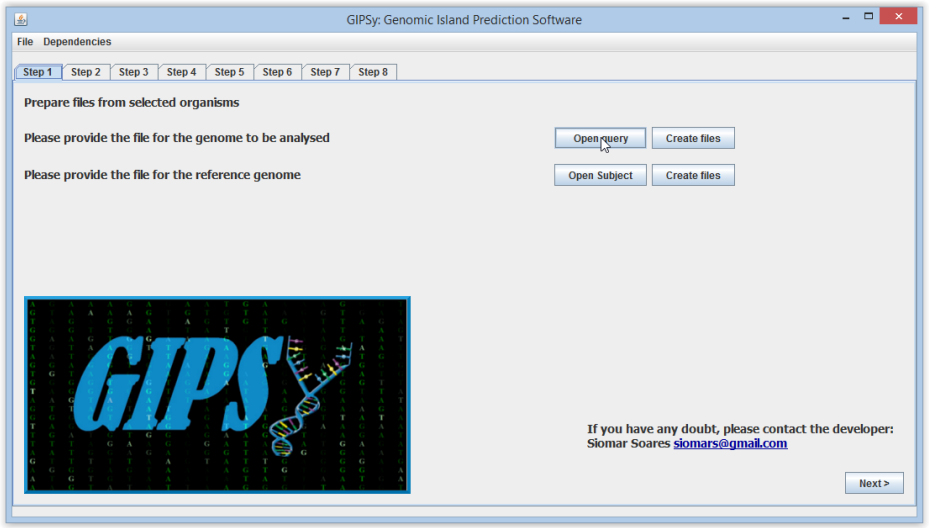
Similarly to the GIPSy.Installer.jar, GIPSy.jar may be run in Windows by simply double clicking, whereas in Linux-based environments, it requires a simple command line, as follow:

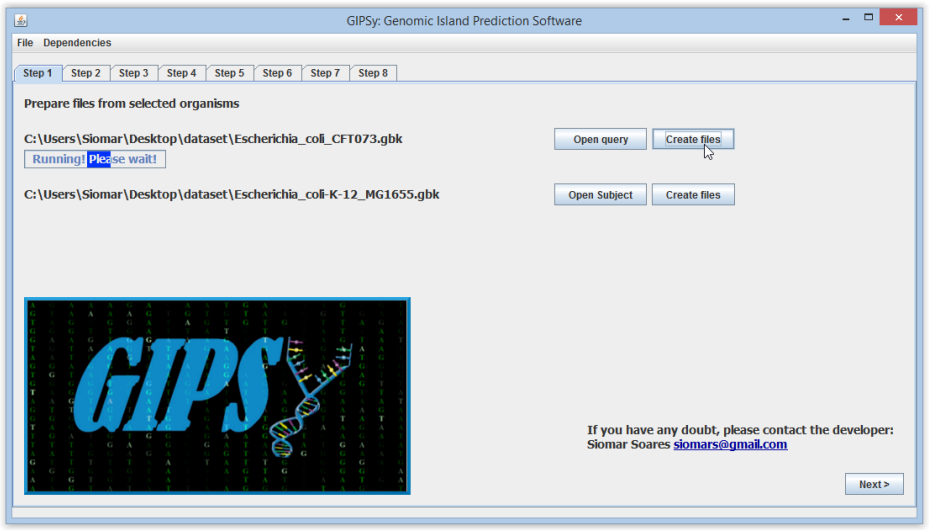
java –jar GIPSy.jar

2.1 Step 1

After the software is open, click the button “Open Query” and choose the file for the pathogenic organism “Escherichia_coli_CFT073.gbk” (downloadable as part of the dataset.zip example files at http://www.bioinformatics.org/downloads/index.php?file_id=587) and click the buttom “Create files”.

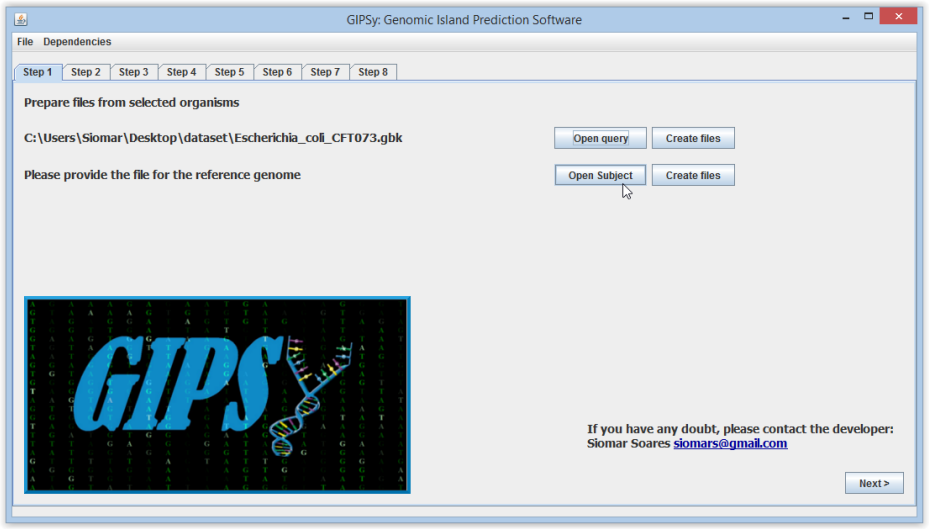
A) Creating files for the query genome.

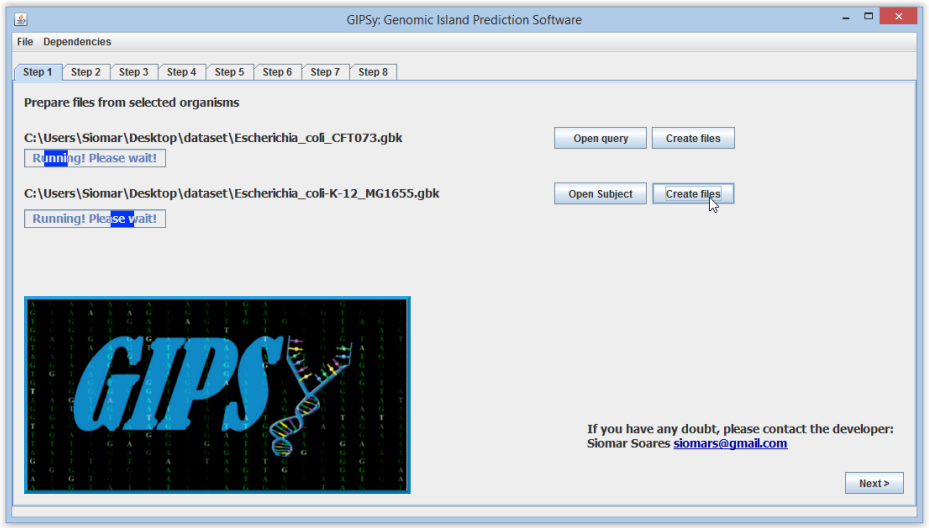




b) Creating files for the subject genome

Click the button “Open Subject” and choose the file for the non-pathogenic closely-related organism “Escherichia_coli-K-12_MG1655.gbk” (downloadable as part of the dataset.zip example files at http://www.bioinformatics.org/downloads/index.php?file_id=587) and click the buttom “Create files”.

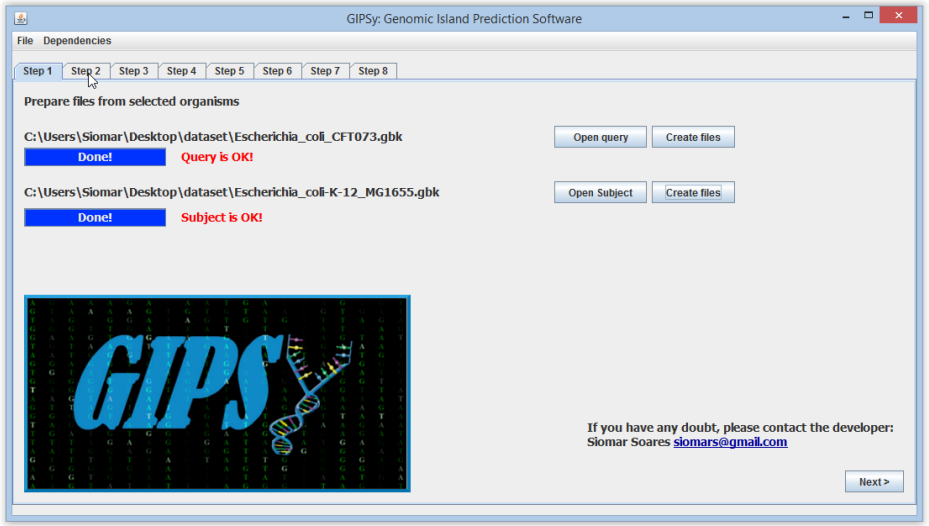


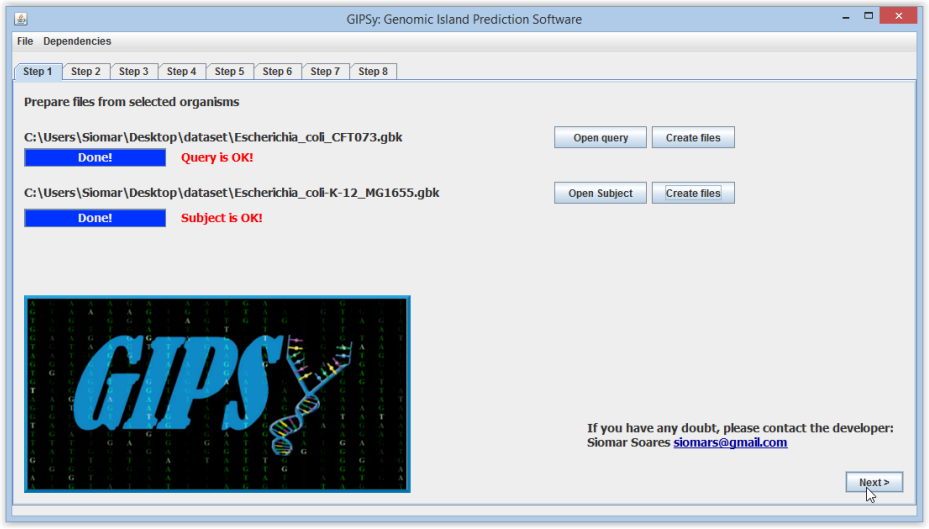


2.2 Going further from one Step to another.

After finishing Step 1, you may click Step 2 or next in order to go further.

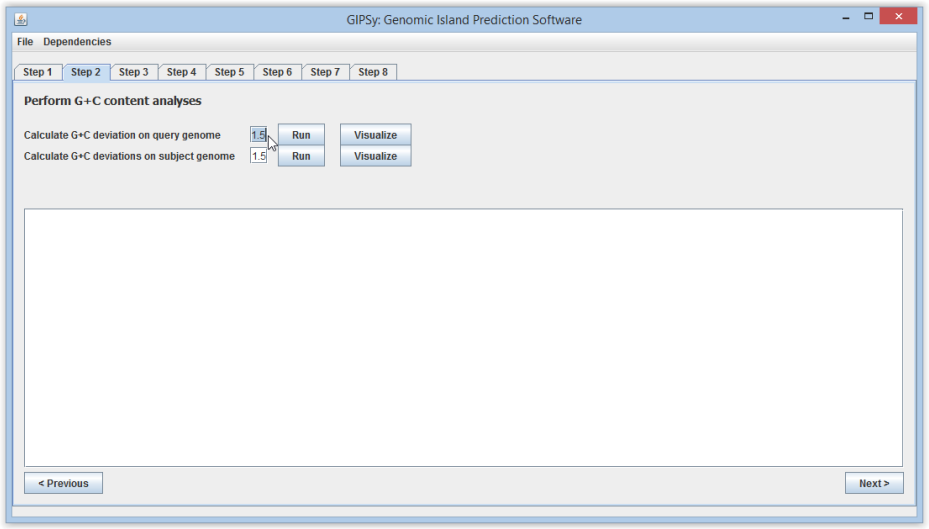
Obs.: All steps are dependent on Step 1. After step 1 is finished, you may run Steps 2-7 concomitantly. Step 8 is dependent on all previous Steps. Be sure to finish all of them before going to Step 8.

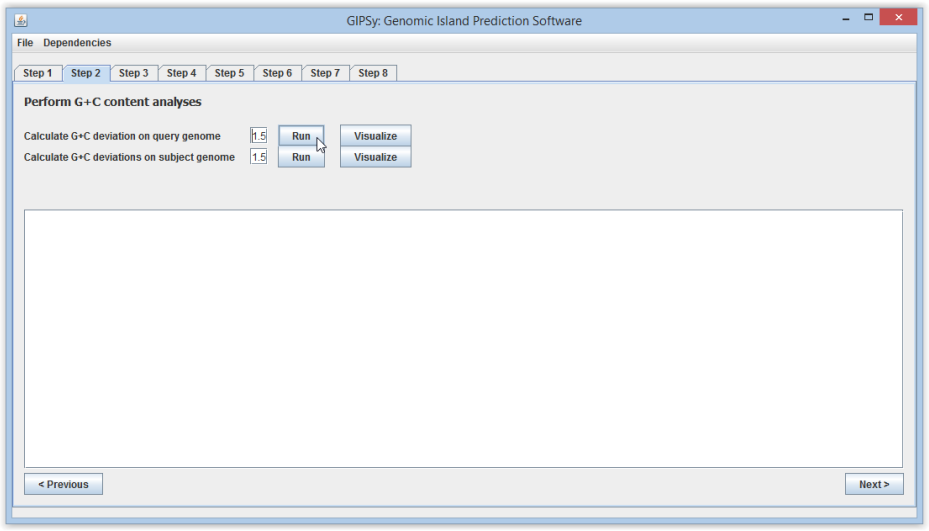


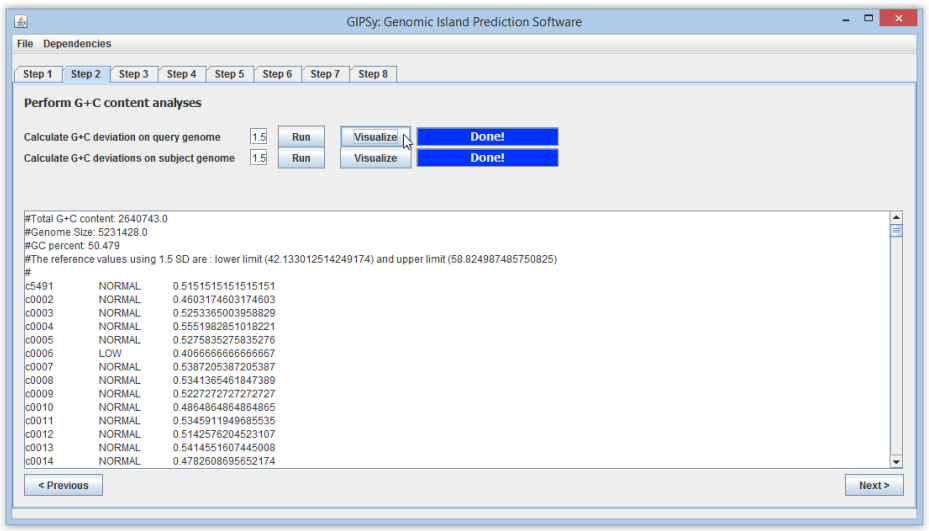


2.3 Step 2

On Step 2, you may choose the cutoff value, in standard deviations (SD), for anomalous G+C analyses. The standard value is 1.5 SDs from the mean. After choosing the cutoff for both query and subject, run the analyses and click visualize in order to see the results.

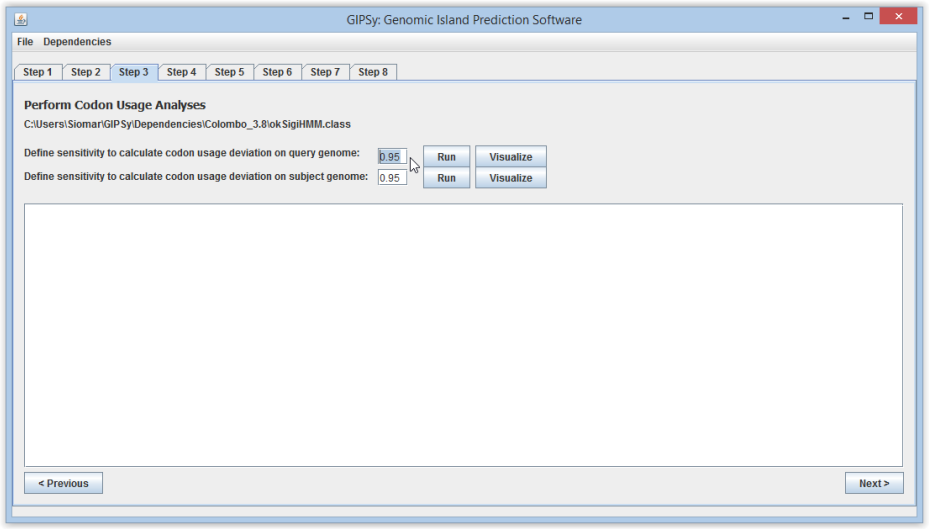


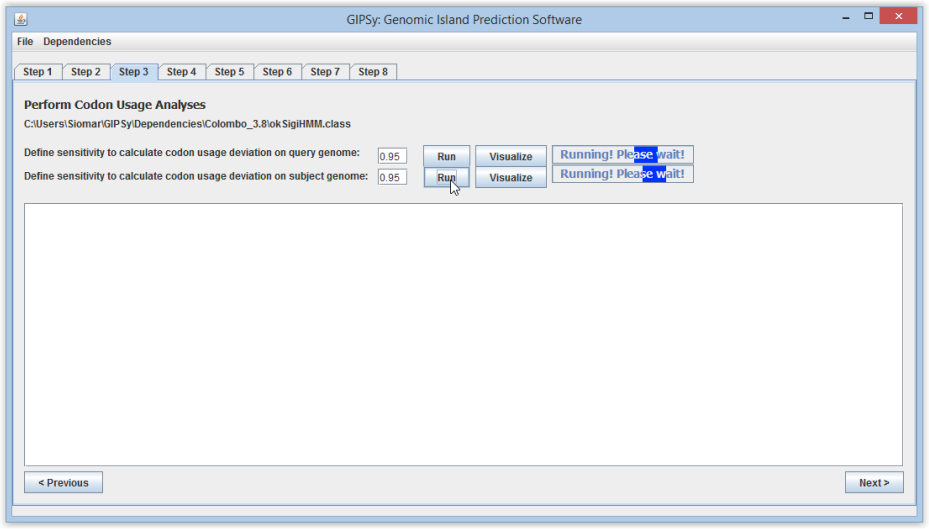


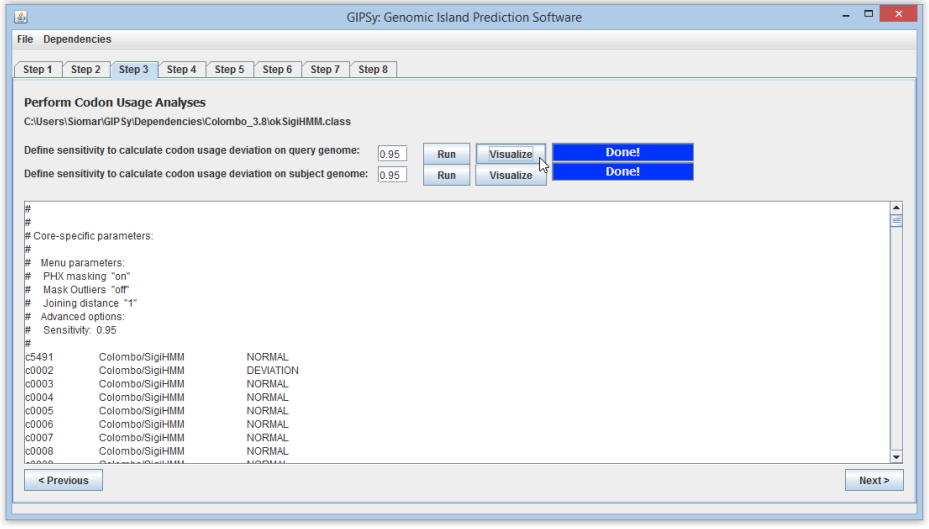


2.4 Step 3

On Step 3, you may choose the sensitivity, from 0.5 to 0.95, for codon usage deviation analyses in Colombo/SigiHMM. The standard value is 0.95. After choosing the cutoff for both query and subject, run the analyses and click visualize in order to see the results.

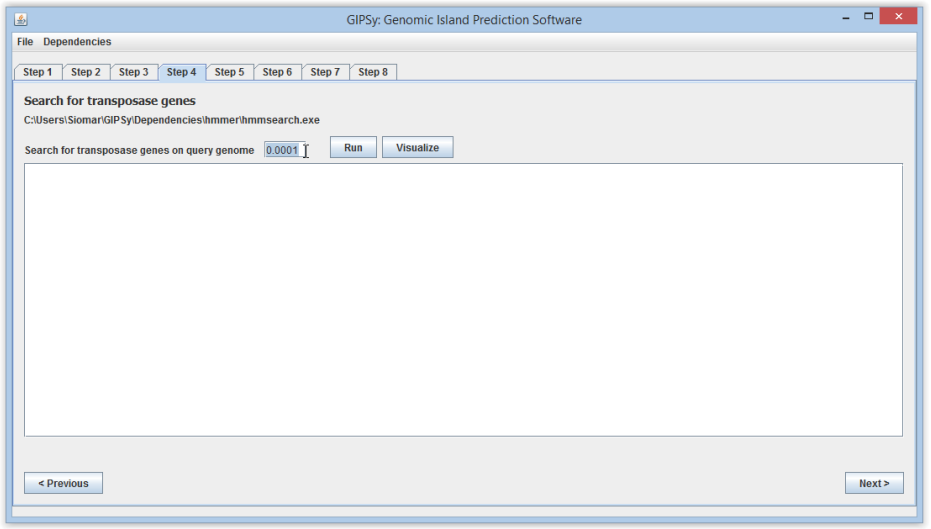


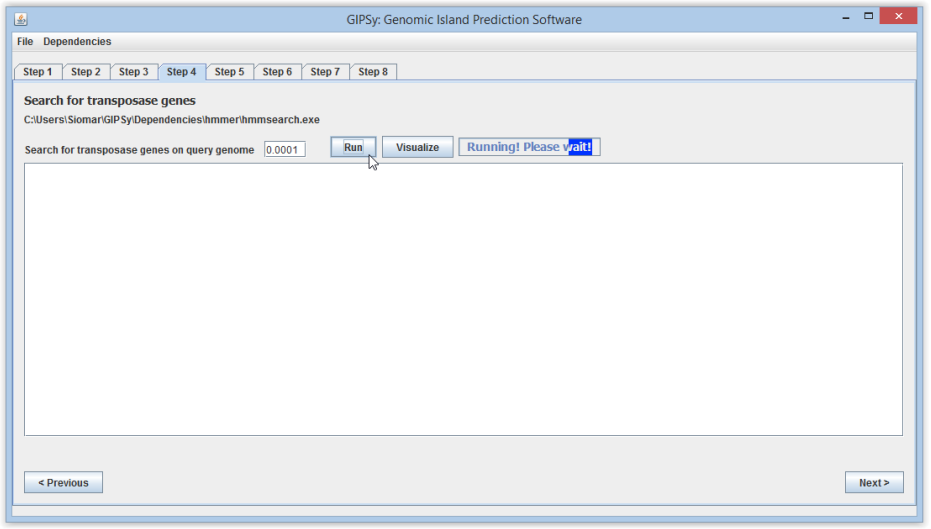


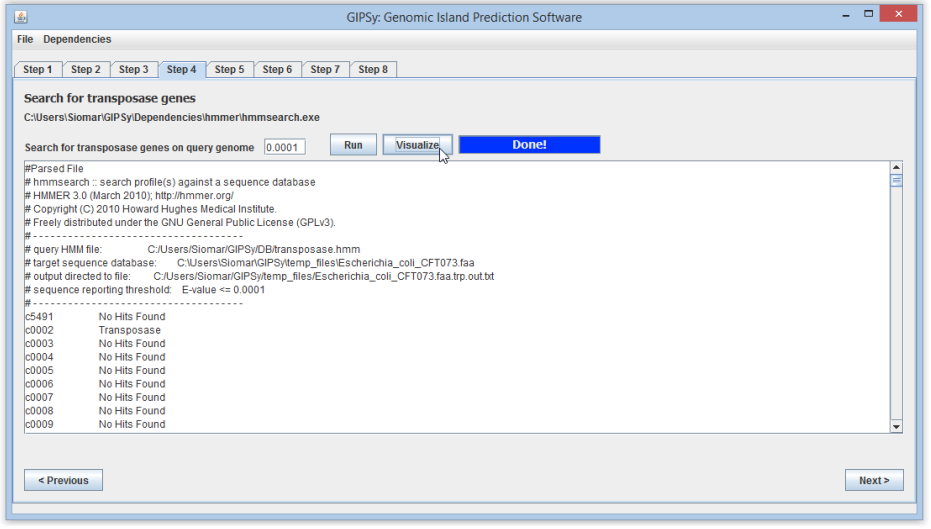


2.5 Step 4

On Step 4, you may choose the e-value for transposase prediction using HMMer. The standard value is 0.0001. After choosing the e-value, run the analyses and click visualize in order to see the results.

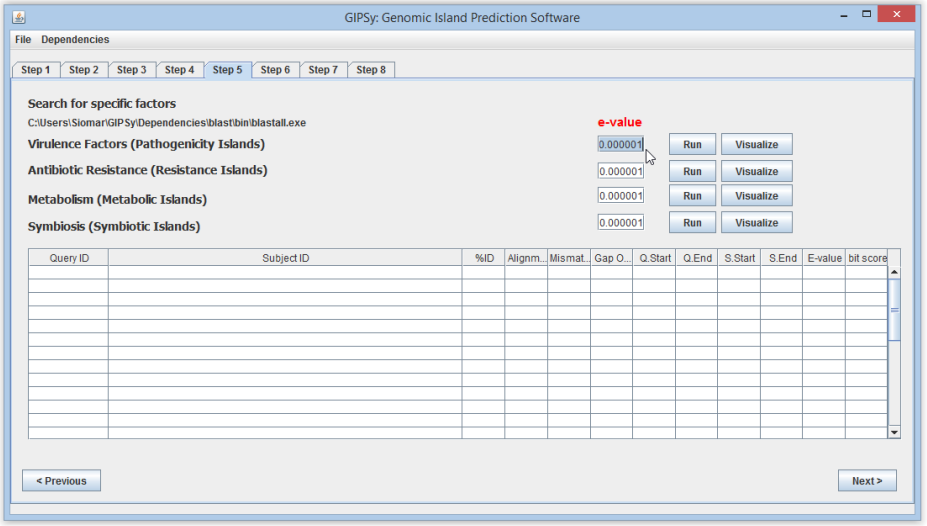


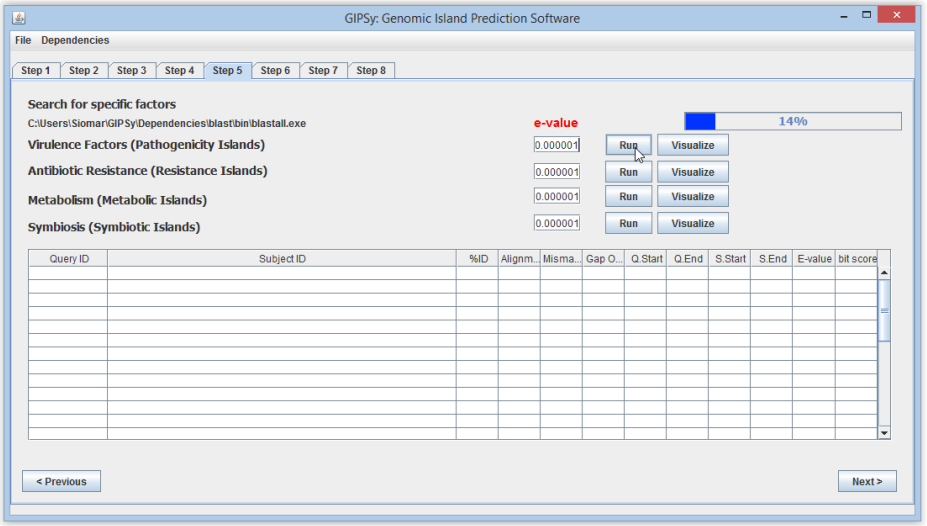


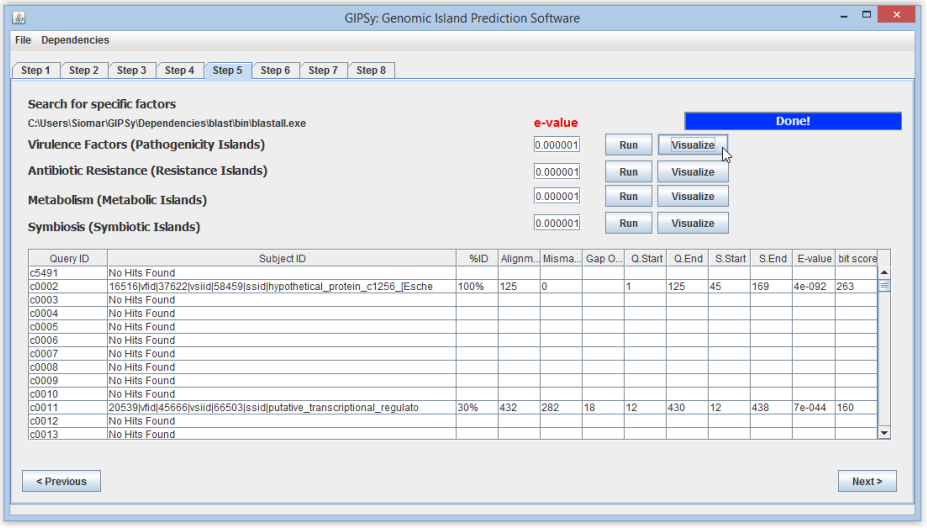


2.6 Step 5

On Step 5, you may choose the e-value for the prediction of specific factors using blastp. The standard value is 0.000001. Moreover, if you are striving to predict Pathogenicity Islands, run the prediction for virulence factors. If you are searching for Resistance Islands, run the analyses for antibiotic resistance genes, and so on. After choosing the e-value, run the analyses for “Virulence Factors” and click visualize in order to see the results.

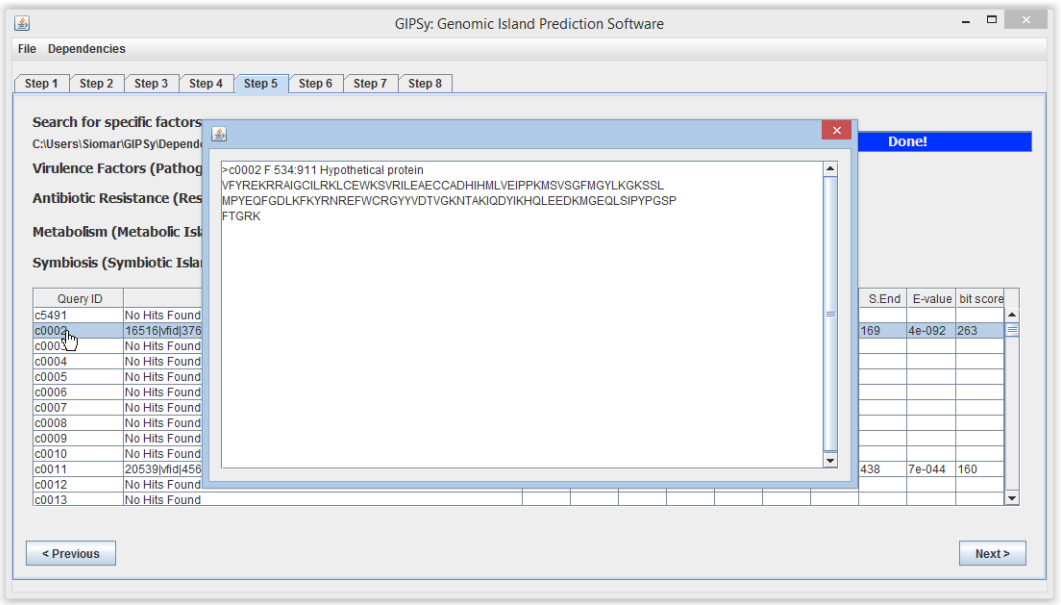




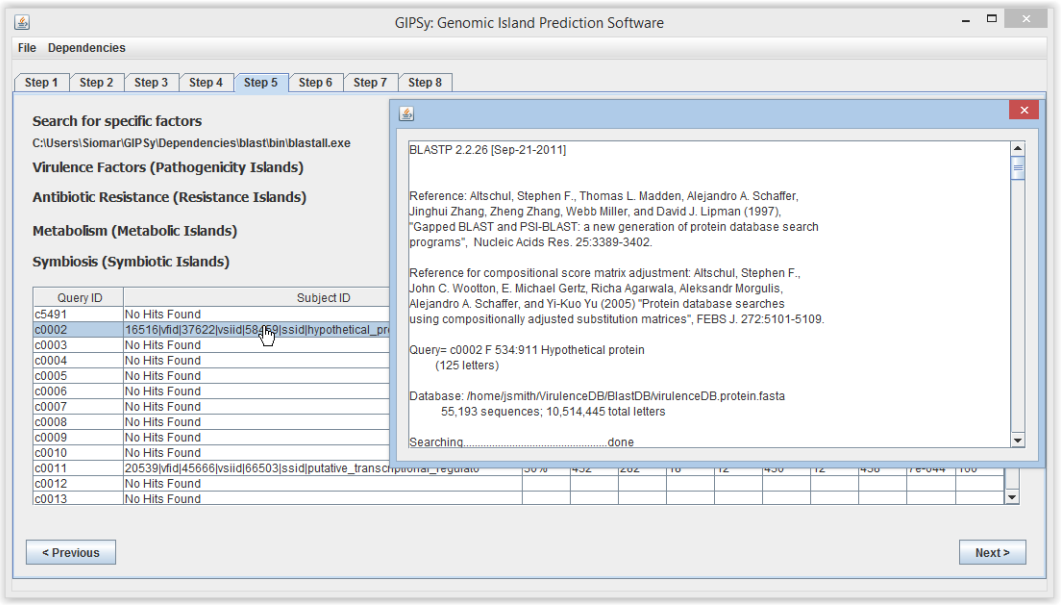


2.6.1 Visualizing the amino acid composition and blast result for each gene (Step 5)

The first and second columns of the table in Step 5 are clickable. Choose some gene and click in the first column in order to see its amino acid composition.

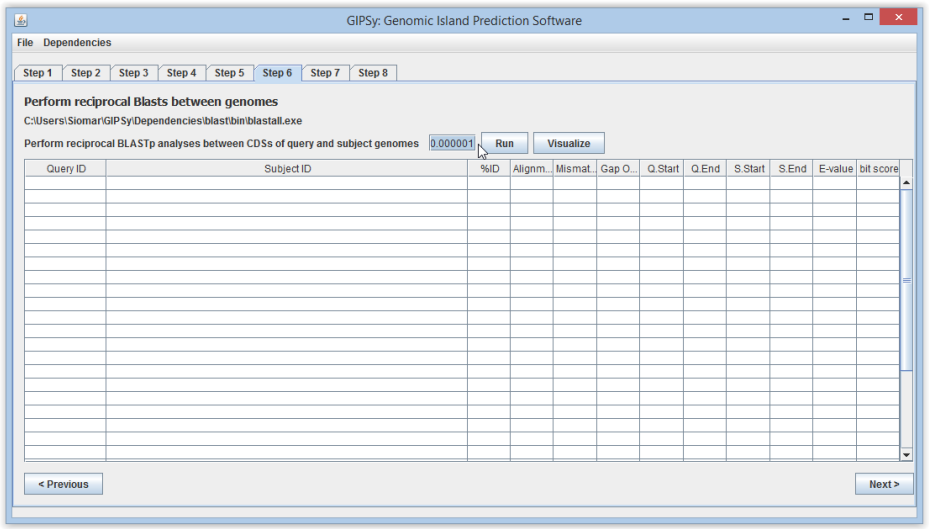


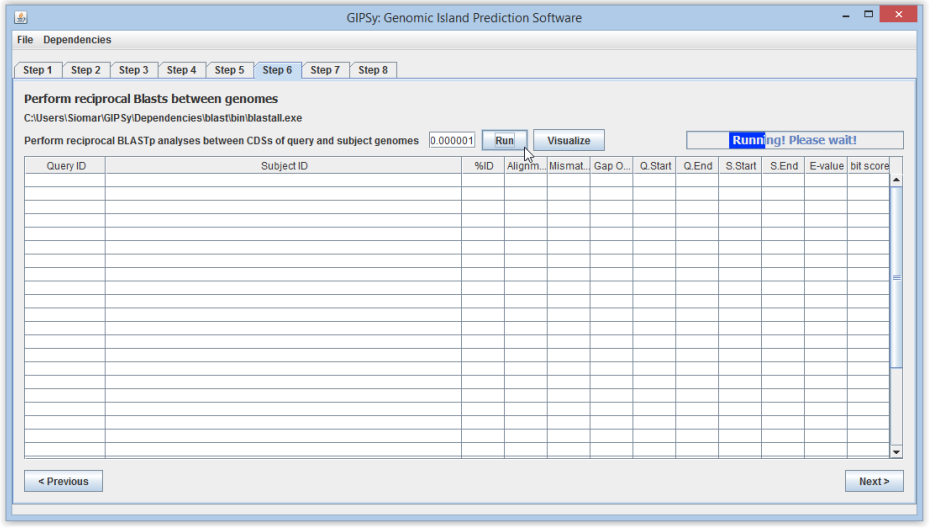
If you want to see the blast result of the given gene against the specific factor database, click the second column and the alignment will appear.

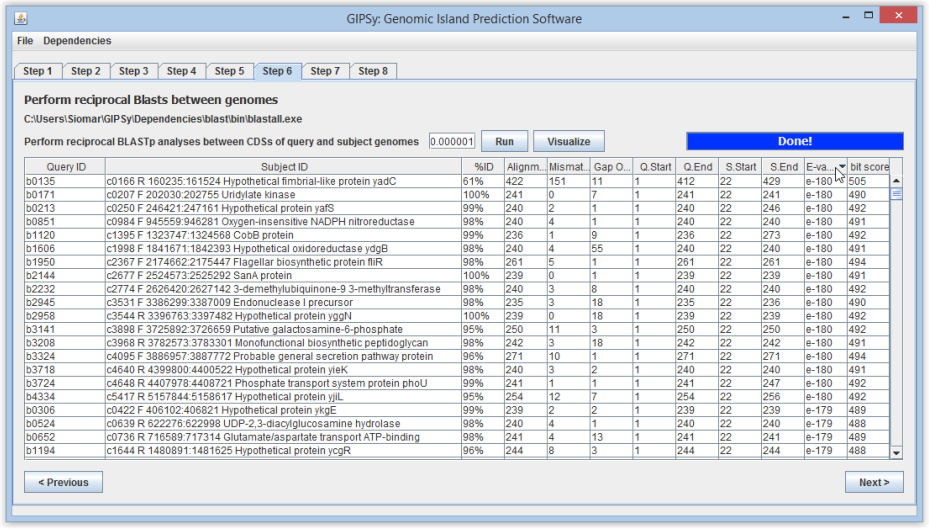


2.7 Step 6

On Step 6, you may choose the e-value for reciprocal blast analyses between the query and subject genomes. The standard value is 0.000001. After choosing the e-value, run the analyses and click visualize in order to see the results.

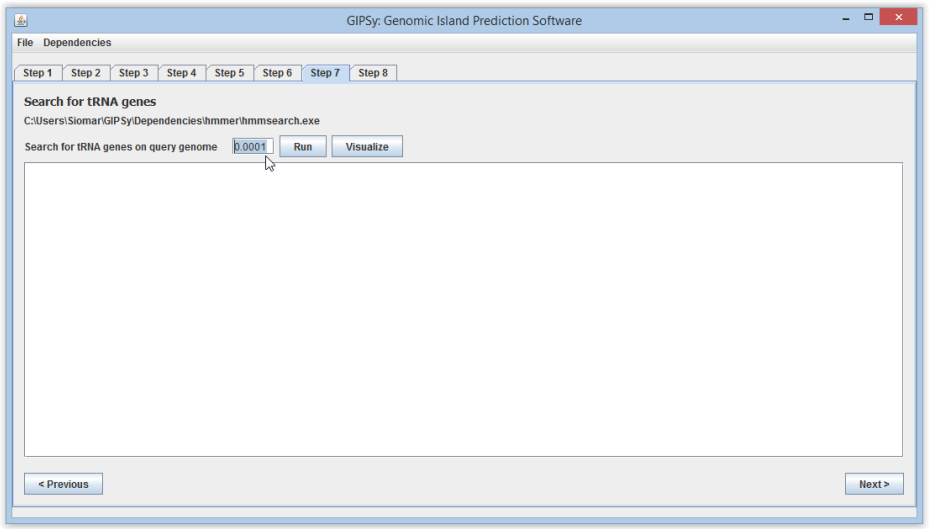


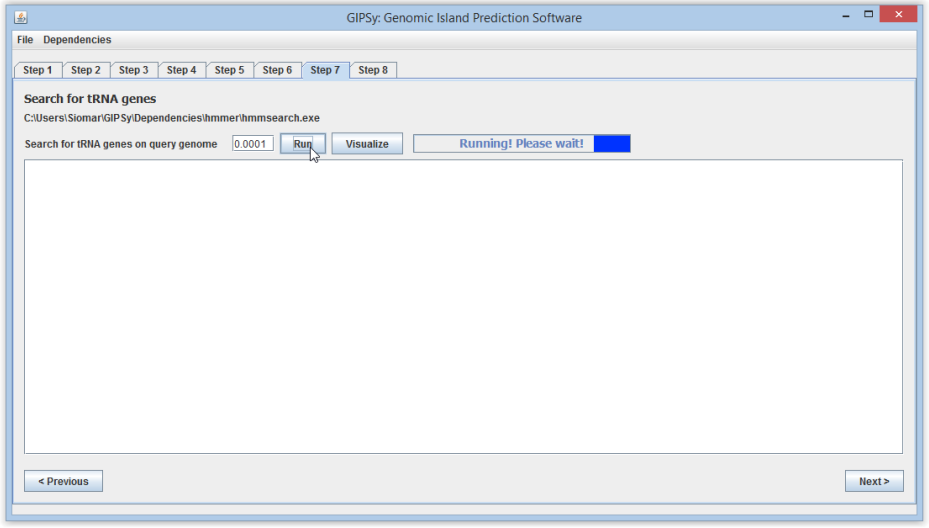


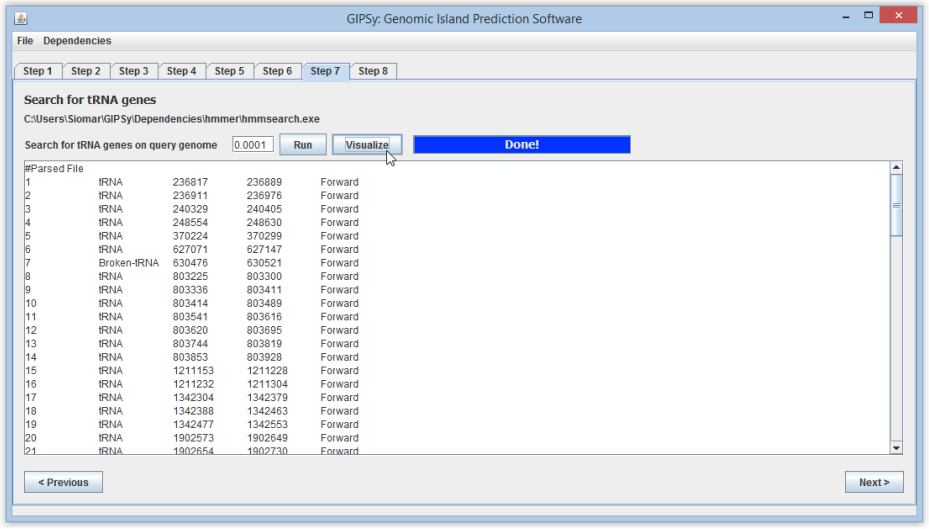


2.8 Step 7

On Step 7, you may choose the e-value for tRNA prediction using HMMer. The standard value is 0.0001. After choosing the e-value, run the analyses and click visualize in order to see the results.

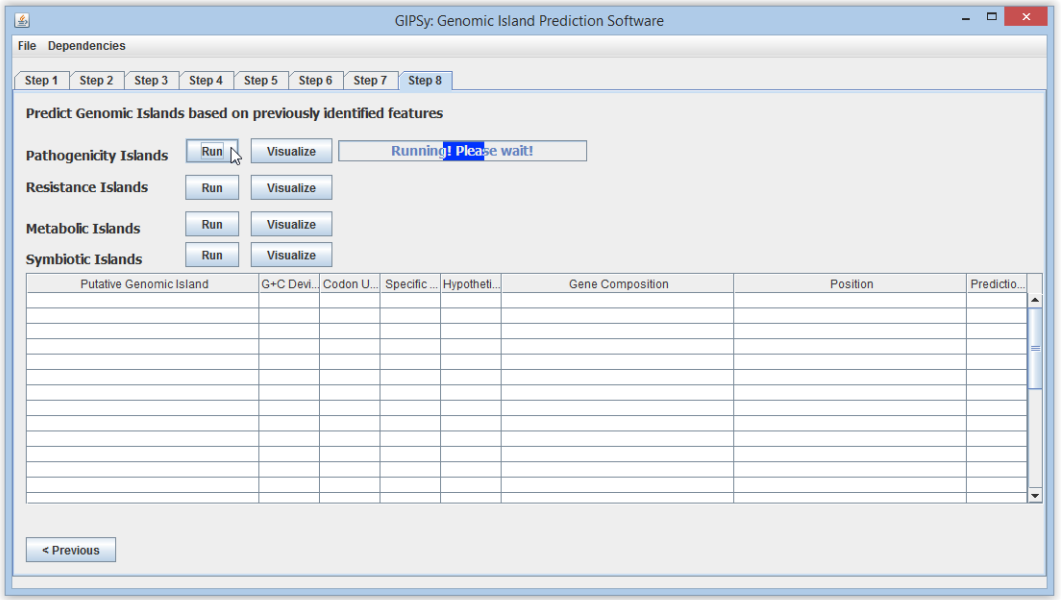


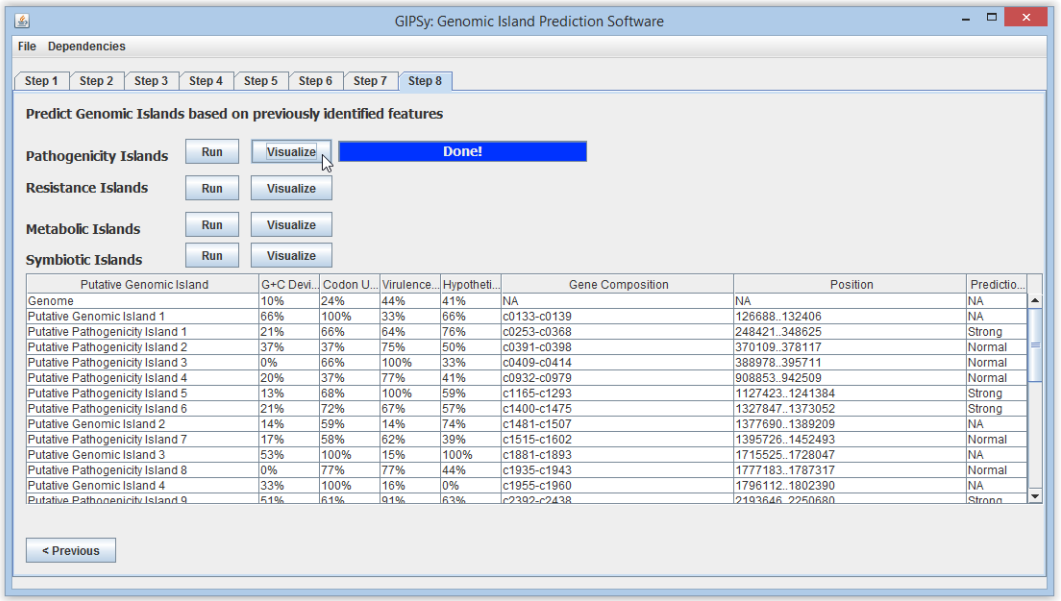




2.9 Step 8

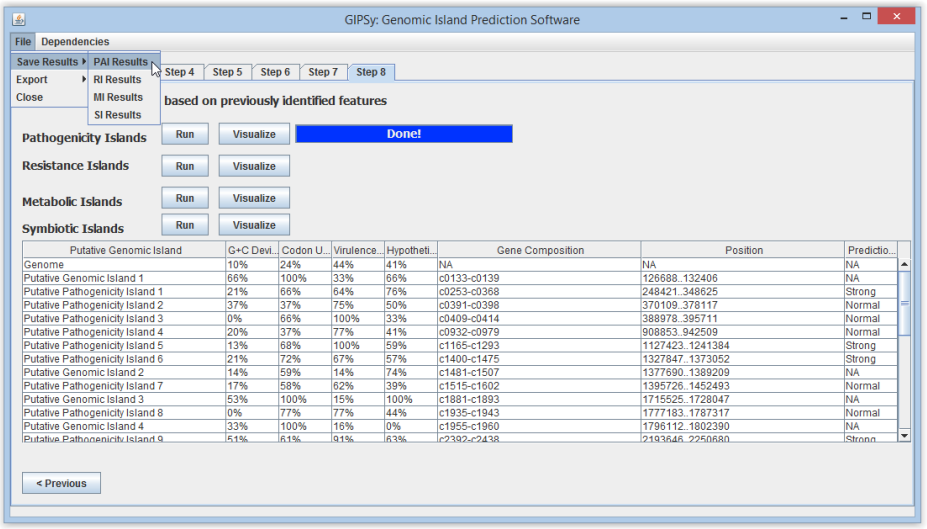
Finally, on Step 8, all the data from the previous steps are used to predict the chosen genomic islands of the given organism. Run the analyses for “Pathogenicity Islands” and click visualize in order to see the results.



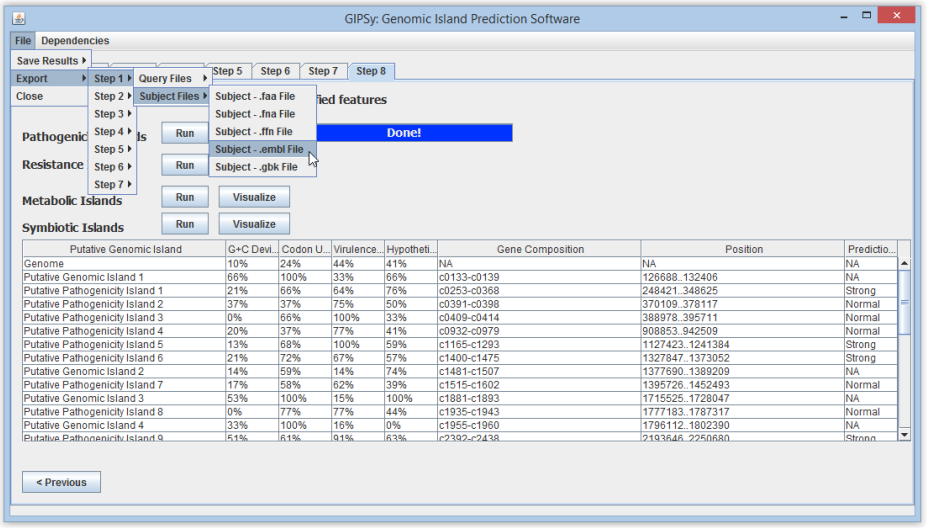


3. Saving results and exporting additional analyses

After all the analyses have been performed, the results may be saved by clicking “File → Save Results → PAI Results”.



Also, the results for each intermediary step may also be exported in “File → Export”.



4. TroubleShooting

4.1 GIPSy breaks at step 5 on Linux

If GIPSy returns an error message at step 5 on Linux, there is probably a limitation on the number of files submitted to blastp. In order to circumvent this problem, please try increasing the ulimit. In Ubuntu, you must modify /etc/security/limits.conf. In a terminal, type:

```
sudo gedit  /etc/security/limits.conf
```

Then, before the end of file, add the following lines:

```
*          soft       nofile     40000
*          hard       nofile     40000
# End of file
```

Also, you need to modify the following just before the end of file:

```
sudo gedit /etc/pam.d/common-session*
```

```
session required pam_limits.so
# end of pam-auth-update config
```

Finally, restart your machine and test your new ulimit typing on the terminal:

```
ulimit -n
```

4.2 GIPSy breaks at steps 4 and 7 on Linux

In this version of GIPSy, we used a 32 bits version of hmmer, which prevents it from working properly on some Linux distributions. We are currently working to update the dependencies package; however, this problem may also be circumvented by installing some 32 bits libraries. On Ubuntu, open a terminal and type:

```
sudo apt-get install libgtk2.0-0:i386 libnss3-1d:i386 libnspr4-0d:i386 lib32nss-mdns libxml2:i386 libxslt1.1:i386 libstdc++6:i386
```

4.3 GIPSy breaks at steps 3 to 7 on Linux

Finally, when the files or their paths have spaces in their names, the third-party software may not work properly. We tried to circumvent this the most as possible. However, the command formatdb from blast is still not working fine. So, in order to circumvent this problem, avoid any space in the names of files.

Wrong: Corynebacterium pseudotuberculosis 1002.gbk

Right: Corynebacterium_pseudotuberculosis_1002.gbk

Also, be sure that the complete path does not have any spaces in names.

Wrong: /home/siomar/gipsy folder/Corynebacterium.gbk

Right: /home/siomar/gipsy_folder/Corynebacterium.gbk

5. Acknowledgments

We would like to kindly thank Taryn Takebayashi for sharing her experience with the software and also for providing a valuable feedback on the problems from sections 4.1 and 4.2.