RepeatAnalyzer Quickstart Guide

Overview

The purpose of this guide is to allow you to set up and begin using the RepeatAnalyzer software package on a new computer, with either Windows, Mac OS X or Linux operating systems. This process involves installing free software from a third party source and thus requires an internet connection. A brief description of how to use the software follows, however this is not meant to be a complete source on use cases. For questions related to problems with installation or usage of this software, contact Helen Catanese at helen.catanese@wsu.edu.

Download

Download the source code and associated data files here.

Direct Download (Recommended)

On the left hand side of the page, click “Downloads.” Then under the Downloads tab, select “Download repository.” Be sure to unzip the directory to a convenient location.

Using Git

If you are familiar with Git, you can also clone this repository by running > git clone https://bitbucket.org/repeatgroup/repeatanalyzer/git

Installation

Before beginning installation, be sure you have an internet connection.

1) Go to www.continuum.io/downloads. This is the download page for the Anaconda Python distribution. Note: while Anaconda is not strictly required to run RepeatAnalyzer, it comes prepackaged with most of the libraries that RepeatAnalyzer needs, and reduces the difficulty of installing those dependencies considerably.

2) The page includes installation instructions for Windows, Mac OS X and Linux. Be sure to follow the installation instructions for your specific platform, and to choose Python version 2.7.

    a) If you are a Mac OS X user, we recommend using the graphical installer, rather than the command line installer.

3) Once the installer has downloaded, run it using the instructions provided on the page below the download link.

    a) If you are using a graphical installer (Mac OS X or Windows), select all the default options. (Confirm that both boxes under “advanced option” are checked.)

    b) If you are using a command line installer (Mac OS X or Linux) after running the installer as directed on the page, be sure to say ‘yes’ when prompted to prepend Anaconda to your PATH.

4) If you are using Linux or Mac OS X, open (or restart) the terminal. If you are using Windows, once you have installed Anaconda, you will have a program called Anaconda Command Prompt, or Anaconda Prompt installed on your machine. Open that.

5) We will use the command prompt to install several additional libraries the program needs. To interact with the command prompt, simply type in a command and hit the 'enter' key. Henceforth, commands to be submitted in this way will be denoted as “> command here”

   a) If you are a Mac OS X or Linux user and at any point you see the error “conda: command not found” or “pip: command not found”, it means you did not add Anaconda to your execution path. To do this add it open the \etc\paths file and add the following line [Path to Anaconda install]/anaconda/bin. To do this, you will need administrator permission.

6) In the Anaconda command prompt, enter the following commands (these are cases sensitive). If an y/n options appear, enter > y

```
> conda install numpy

> conda install basemap

> pip install pyshp

> pip install geopy
```

7) Be sure that you have downloaded the source code directory from the link above, and if necessary, unzipped the file. The files in this folder must remain together in order to function properly. Do not remove any of them from this folder (or move them to subfolders).

8) Once this is completed, if you are using Windows simply double-click the RepeatAnalyzer.py file in the source code folder to open the program. If you are using a Mac or Linux, navigate into the source folder in the terminal and run the command “python Repeatanayzer.py” to start the program. You may create a shortcut to this file to run the program from a more convenient location.

   a) If you are prompted to choose a program to open the file with, navigate to the Anaconda2 installation directory, and select the python.exe (NOT pythonw.exe) file.

      i) In rare cases on Windows the python.exe file will not be selectable as a default program. To correct this, run the following commands in the windows command prompt (cmd.exe). Be sure to run it as an administrator.

```
> assoc .py=Python.File

> ftype Python.File=“[Anaconda installation path]\python.exe” “%1” %*
```

   b) If the window opens, but closes immediately, there is likely a problem with your installation, or a conflict with other software installed on your system (like other versions of Python). In this case, see troubleshooting.

9) When the program is run for the first time, it will warn you that the RepeatAnalyzer.dat file is missing, and ask if you wish to continue. This is normal behavior. To setup the program with the default *A. marginale* data included with the source code, just enter ‘y’ and then ‘Anaplasma marginale’ as the species name. See ‘Using the Main Functions’ subsection 5 for details on loading in the data for the first time.

Basics

RepeatAnalyzer uses a primarily command line interface, which means that interacting with the main menu of the program will usually involve reading options from the screen and typing in your selection. The main menu includes 11 options (currently) which can be selected by entering the corresponding number for that option. Once an option has been selected, RepeatAnalyzer may open additional dialogs, or present other command line options.

At any given time, RepeatAnalyzer will be set to work on a single species (*Anaplama Marginale*, *Anaplasma Centrale*, etc). Be sure that the program is set to the intended species when you are adding or searching data, as it will only access the repeats and strains stored for that particular species. If the species you wish to work with does not appear as on option in command 2 (change current species), you may add it with command 2, sub-option 0. Be sure to check the current species each time you open the program (if you are storing data for more than one species).

Adding Data

With existing data

If the RepeatAnalyzer.dat file is in the same folder as RepeatAnalyzer.py when it is executed, the stored data will be loaded. Upon opening the program you will see the main menu, showing the current species, and the number of strains and repeats stored respectively. If this is the correct species, continue as usual, otherwise, be sure to change the current species before moving on.

Without existing data

If the RepeatAnalyzer.dat file is not in the same folder as the RepeatAnalyzer.py file when it is run, the program will first confirm that the RepeatAnalyzer.dat file is missing and ask if you wish to continue. If you enter 'y' (for yes) it will ask what species you would like to work with. At this point you will be prompted to enter the name (Genus species) of the species you will be adding data for. After you have done this the main menu of the program will appear, with the species name you entered and 0 strains, 0 repeats.

In either case...

At the main menu, select command 5 and enter the name of the input file. AmarginaleData.txt comes packaged with the source code. If you wish to use another input file, make sure it follows the formatting outlined in the Sample File Formatting Section below. The program will then read this file into its internal storage. If there is an issue reading, the program will try to give the line number where the issue occurred in the file, but the error may actually be on the line before or after the one given, so check that whole area and correct any formatting mistakes until it reads in with no errors. Once reading is complete, you will be prompted on whether you would like to update the geocoding for your data. If you have an internet connection, select ‘y’, as mapping and regional analysis functionalities will be disabled until this is completed. If you select ‘n’ now, geocoding can be updated later via command 11.

Keep in mind that any strains or repeats with the same sequence will be stored as a single repeat or strain with multiple names, so if the number of repeats is lower than you were expecting after adding a file, this is likely the reason. Check the data summary (command 10) for details on where this occurred.

Note that within the included Anaplasma marginale data file, the repeat ‘tc63_3_s06’ is present with no given sequence. This is not a mistake, as it was included in a publication with no sequence provided and no valid reference. It will not cause any problems in the program, only an error message when inputting the file; genotypes including it will be ignored. Keep in mind that this is the 235th repeat, as referenced in the main paper.

Sample Input File Format

```
Paper:
        [citation line 1: title]
                [citation line 2:autors]
                [citation line 3: other info]
        Repeats:
                [name] : [sequence]
                {repeat for each repeat}
        Strains:
                [name (optional, but the colon is not) OR year , animalID (if using auto-naming)] :
                [space separated list of repeat names] : [location in the form country,province,county
                or country province or country]
                {repeat this for each strain. if you have one strain at multiple locations it will need to be
                listed twice, but rest assured it will only be counted once}
{Repeat this pattern for each published paper}

Unpublished:
        [author name(s), this should look like 2 of paper citation]
                [year of findings]
        Repeats:
                {this should be exactly the same as the repeats section for a paper}
        Strains:
                {likewise, this is the same as the strains section for a paper}
{Repeat this pattern for each unpublished source}
```

Using the Main Functions

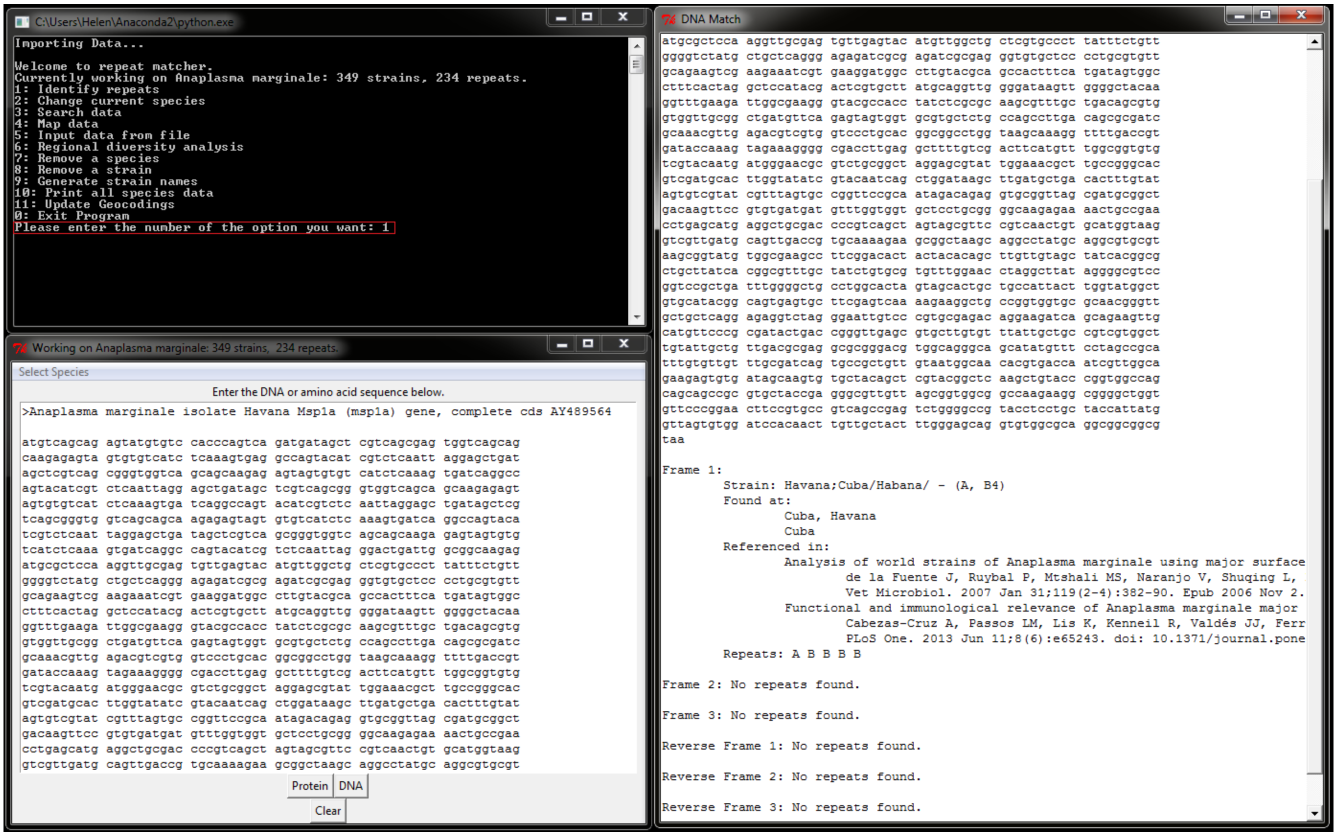
Each of the following subsections outlines the capabilities of one of the 11 main menu functions which RepeatAnalyzer performs. Each section includes details on any input required for the function, and any output it generates. A user can access the function by entering the associated number at the main menu.

1. Identify repeats

Input: One full or partial gene sequence either in DNA or protein form

OR multiple gene sequences in FASTA format.

Output: The names of the maximal set of repeats in each sample with genotype name and publication details if any.


C:\Users\Helen\Anaconda2\python.exe
Working on Anaplasma marginale: 349 strains, 234 repeats.
Select Species
Enter the DNA or amino acid sequence below.
Protein
DNA
Clear
DNA Match

```
Importing Data...

Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 1
```

```
>Anaplasma marginale isolate Havana Msp1a (msp1a) gene, complete cds AY489564

atgtcagcag agtatgtgtc cacccagtca gatgatagct cgtcagcgag tggtcagcag
caagagagta gtgtgtcatc tcaaagtgag gccagtacat cgtctcaatt aggagctgat
agctcgtcag cgggtggtca gcagcaagag agtagtgtgt catctcaaag tgatcaggcc
agtacatcgt ctcaattagg agctgatagc tcgtcagcgg gtggtcagca gcaagagagt
agtgtgtcat ctcaaagtga tcaggccagt acatcgtctc aattaggagc tgatagctcg
tcagcgggtg gtcagcagca agagagtagt gtgtcatctc aaagtgatca ggccagtaca
tcgtctcaat taggagctga tagctcgtca gcgggtggtc agcagcaaga gagtagtgtg
tcatctcaaa gtgatcaggc cagtacatcg tctcaattag ggactgattg gcggcaagag
atgcgctcca aggttgcgag tgttgagtac atgttggctg ctcgtgccct tatttctgtt
ggggtctatg ctgctcaggg agagatcgcg agatcgcgag ggtgtgctcc cctgcgtgtt
gcagaagtcg aagaaatcgt gaaggatggc cttgtacgca gccactttca tgatagtggc
ctttcactag gctccatacg actcgtgctt atgcaggttg gggataagtt ggggctacaa
ggtttgaaga ttggcgaagg gtacgccacc tatctcgcgc aagcgtttgc tgacagcgtg
gtggttgcgg ctgatgttca gagtagtggt gcgtgctctg ccagccttga cagcgcgatc
gcaaacgttg agacgtcgtg gtccctgcac ggcggcctgg taagcaaagg ttttgaccgt
gataccaaag tagaaagggg cgaccttgag gcttttgtcg acttcatgtt tggcggtgtg
tcgtacaatg atgggaacgc gtctgcggct aggagcgtat tggaaacgct tgccgggcac
gtcgatgcac ttggtatatc gtacaatcag ctggataagc ttgatgctga cactttgtat
agtgtcgtat cgtttagtgc cggttccgca atagacagag gtgcggttag cgatgcggct
gacaagttcc gtgtgatgat gtttggtggt gctcctgcgg ggcaagagaa aactgccgaa
cctgagcatg aggctgcgac cccgtcagct agtagcgttc cgtcaactgt gcatggtaag
gtcgttgatg cagttgaccg tgcaaaagaa gcggctaagc aggcctatgc aggcgtgcgt
```

```
atgcgctcca aggttgcgag tgttgagtac atgttggctg ctcgtgccct tatttctgtt
ggggtctatg ctgctcaggg agagatcgcg agatcgcgag ggtgtgctcc cctgcgtgtt
gcagaagtcg aagaaatcgt gaaggatggc cttgtacgca gccactttca tgatagtggc
ctttcactag gctccatacg actcgtgctt atgcaggttg gggataagtt ggggctacaa
ggtttgaaga ttggcgaagg gtacgccacc tatctcgcgc aagcgtttgc tgacagcgtg
gtggttgcgg ctgatgttca gagtagtggt gcgtgctctg ccagccttga cagcgcgatc
gcaaacgttg agacgtcgtg gtccctgcac ggcggcctgg taagcaaagg ttttgaccgt
gataccaaag tagaaagggg cgaccttgag gcttttgtcg acttcatgtt tggcggtgtg
tcgtacaatg atgggaacgc gtctgcggct aggagcgtat tggaaacgct tgccgggcac
gtcgatgcac ttggtatatc gtacaatcag ctggataagc ttgatgctga cactttgtat
agtgtcgtat cgtttagtgc cggttccgca atagacagag gtgcggttag cgatgcggct
gacaagttcc gtgtgatgat gtttggtggt gctcctgcgg ggcaagagaa aactgccgaa
cctgagcatg aggctgcgac cccgtcagct agtagcgttc cgtcaactgt gcatggtaag
gtcgttgatg cagttgaccg tgcaaaagaa gcggctaagc aggcctatgc aggcgtgcgt
aagcggtatg tggcgaagcc ttcggacact actacacagc ttgttgtagc tatcacggcg
ctgcttatca cggcgtttgc tatctgtgcg tgtttggaac ctaggcttat aggggcgtcc
ggtccgctga tttggggctg cctggcacta gtagcactgc tgccattact tggtatggct
gtgcatacgg cagtgagtgc ttcgagtcaa aagaaggctg ccggtggtgc gcaacgggtt
gctgctcagg agaggtctag ggaattgtcc cgtgcgagac aggaagatca gcagaagttg
catgttcccg cgatactgac cgggttgagc gtgcttgtgt ttattgctgc cgtcgtggct
tgtattgctg ttgacgcgag gcgcgggacg tggcagggca gcatatgttt cctagccgca
tttgtgttgt ttgcgatcag tgccgctgtt gtaatggcaa cacgtgacca atcgttggca
gaagagtgtg atagcaagtg tgctacagct cgtacggctc aagctgtacc cggtggccag
cagcagccgc gtgctaccga gggcgttgtt agcggtggcg gccaagaagg cggggctggt
gttcccggaa cttccgtgcc gtcagccgag tctggggccg tacctcctgc taccattatg
gttagtgtgg atccacaact tgttgctact ttgggagcag gtgtggcgca ggcggcggcg
taa

Frame 1:
        Strain: Havana;Cuba/Habana/ - (A, B4)
        Found at:
                Cuba, Havana
                Cuba
        Referenced in:
                Analysis of world strains of Anaplasma marginale using major surface
                        de la Fuente J, Ruybal P, Mtshali MS, Naranjo V, Shuqing L,
                        Vet Microbiol. 2007 Jan 31;119(2-4):382-90. Epub 2006 Nov 2.
                Functional and immunological relevance of Anaplasma marginale major
                        Cabezas-Cruz A, Passos LM, Lis K, Kenneil R, Valdés JJ, Ferr
                        PLoS One. 2013 Jun 11;8(6):e65243. doi: 10.1371/journal.pone
        Repeats: A B B B B

Frame 2: No repeats found.

Frame 3: No repeats found.

Reverse Frame 1: No repeats found.

Reverse Frame 2: No repeats found.

Reverse Frame 3: No repeats found.
```

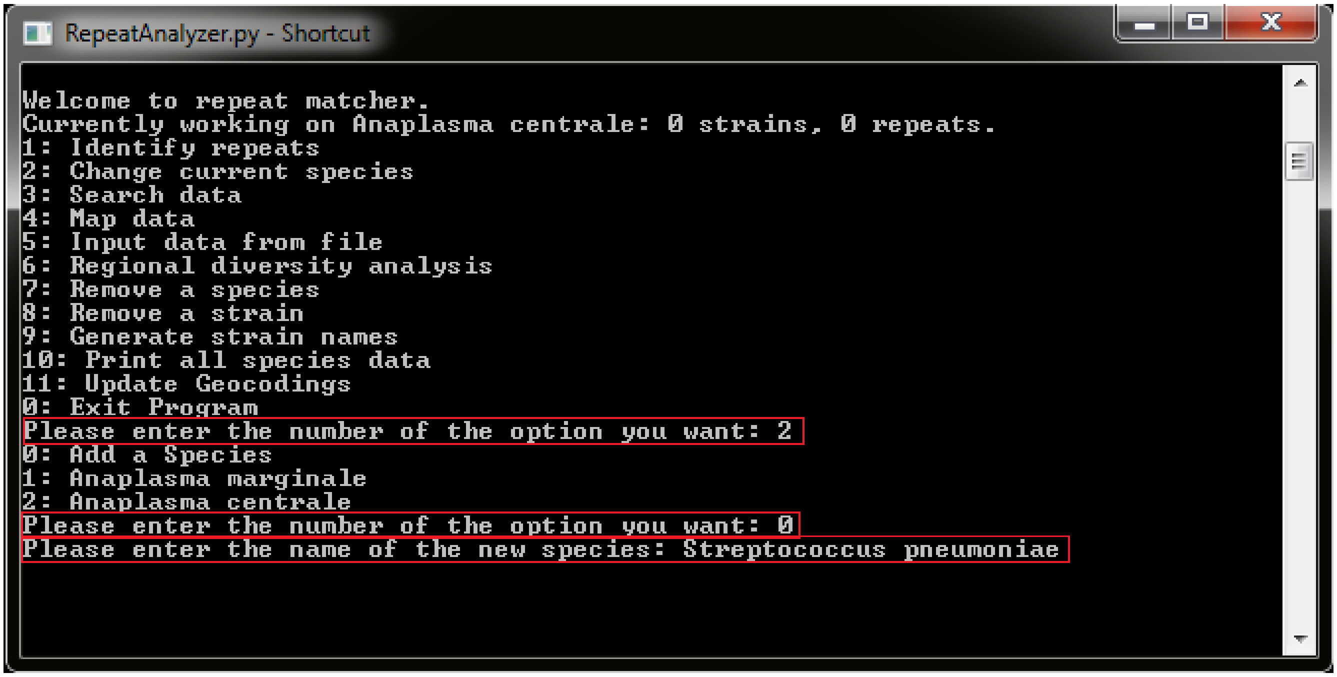
Figure 1 Shows the repeat/genotype identification interface. Once a sequence or sequences have been entered into the identification window (bottom left) and the correct input type is chosen (DNA or Protein), the match window will appear with which (known) repeats occur in the sequence and all relevant information on the genotype (if it has been reported previously). All windows can be resized and/or have their contents copied as needed.

2. Change current species

Input: The number of the species (as listed on-screen)

OR 0 followed by name of the new species.

Output: The main menu header will change as appropriate.



```
Welcome to repeat matcher.
Currently working on Anaplasma centrale: 0 strains, 0 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 2
0: Add a Species
1: Anaplasma marginale
2: Anaplasma centrale
Please enter the number of the option you want: 0
Please enter the name of the new species: Streptococcus pneumoniae
```

Figure 2 Shows the interface to change species or add a new species. Specifically, showing the series of commands used to add a new species to RepeatAnalyzer. The first command opens the species change menu. The second selects add a species. Finally, the third accepts the name of the species from the user. Before the species is added, RepeatAnalyzer will also ask the user to enter the command 'y' to confirm that the new name is correct.

3. Search data

Input: Select entity type by number, repeat, strain or location. Check boxes as desired.

Repeat: repeat name or sequence

OR multiple of repeat names or sequences in FASTA format

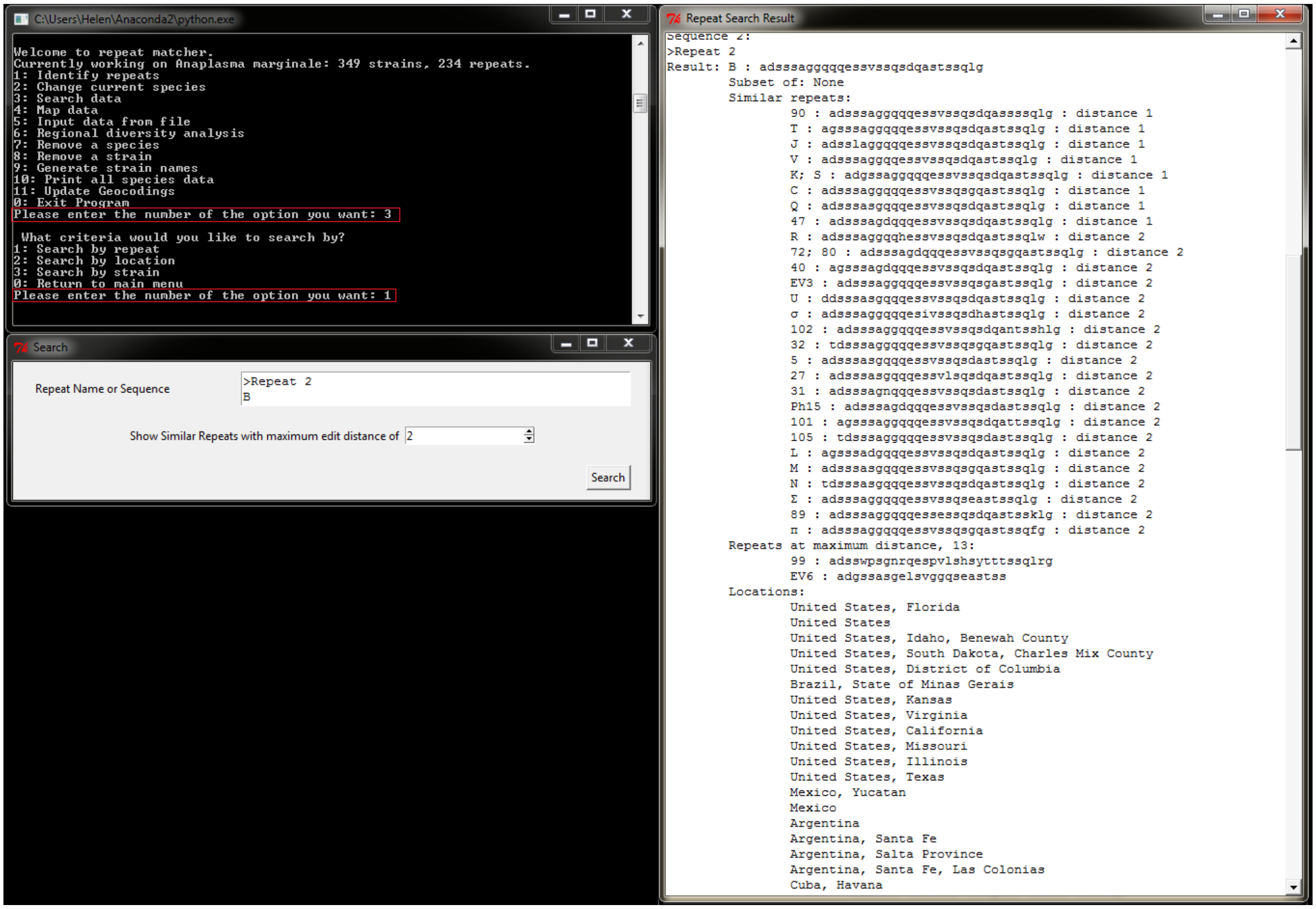
Strain: genotype name or sequence (by repeat names)

OR multiple genotype names or sequences in FASTA format

Location: location name from dropdown

Output: A summary of information for the entity or entities selected.

See Figures 3, 4 and 5 for illustrations of each of the three search types.



C:\Users\Helen\Anaconda2\python.exe

```
Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 3

 What criteria would you like to search by?
1: Search by repeat
2: Search by location
3: Search by strain
0: Return to main menu
Please enter the number of the option you want: 1
```

Search

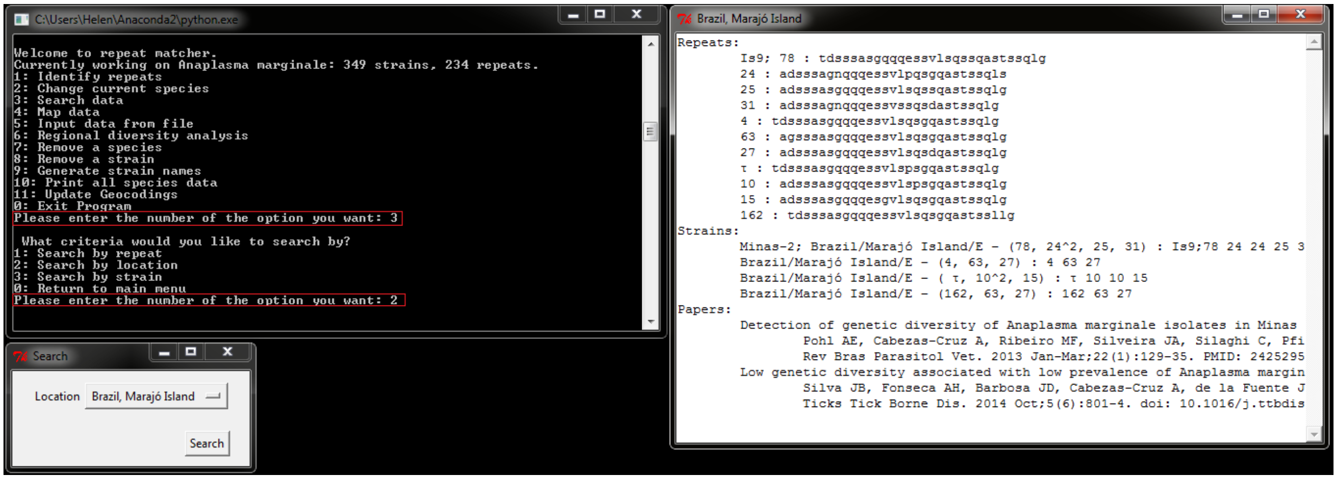
Repeat Name or Sequence: >Repeat 2
B

Show Similar Repeats with maximum edit distance of: 2

Repeat Search Result

```
Sequence 2:
>Repeat 2
Result: B : adsssaggqqqessvssqsdqastssqlg
        Subset of: None
        Similar repeats:
                90 : adsssaggqqqessvssqsdqasssssqlg : distance 1
                T : agsssaggqqqessvssqsdqastssqlg : distance 1
                J : adsslaggqqqessvssqsdqastssqlg : distance 1
                V : adsssaggqqessvssqsdqastssqlg : distance 1
                K; S : adgssaggqqqessvssqsdqastssqlg : distance 1
                C : adsssaggqqqessvssqsgqastssqlg : distance 1
                Q : adsssasgqqqessvssqsdqastssqlg : distance 1
                47 : adsssagdqqqessvssqsdqastssqlg : distance 1
                R : adsssaggqqhessvssqsdqastssqlw : distance 2
                72; 80 : adsssagdqqqessvssqsgqastssqlg : distance 2
                40 : agsssagdqqqessvssqsdqastssqlg : distance 2
                EV3 : adsssaggqqqessvssqsgastssqlg : distance 2
                U : ddsssasgqqqessvssqsdqastssqlg : distance 2
                σ : adsssaggqqqesivssqsdhastssqlg : distance 2
                102 : adsssaggqqqessvssqsdqantsshlg : distance 2
                32 : tdsssaggqqqessvssqsgqastssqlg : distance 2
                5 : adsssasgqqqessvssqsdastssqlg : distance 2
                27 : adsssasgqqqessvlsqsdqastssqlg : distance 2
                31 : adsssagnqqqessvssqsdastssqlg : distance 2
                Ph15 : adsssagdqqqessvssqsdastssqlg : distance 2
                101 : agsssaggqqqessvssqsdqattssqlg : distance 2
                105 : tdsssaggqqqessvssqsdastssqlg : distance 2
                L : agsssadgqqqessvssqsdqastssqlg : distance 2
                M : adsssasgqqqessvssqsgqastssqlg : distance 2
                N : tdsssasgqqqessvssqsdqastssqlg : distance 2
                Σ : adsssaggqqqessvssqseastssqlg : distance 2
                89 : adsssaggqqqessessqsdqastssklg : distance 2
                π : adsssaggqqqessvssqsgqastssqfg : distance 2
        Repeats at maximum distance, 13:
                99 : adsswpsgnrqespvlshsytttssqlrg
                EV6 : adgssasgelsvggqseastss
        Locations:
                United States, Florida
                United States
                United States, Idaho, Benewah County
                United States, South Dakota, Charles Mix County
                United States, District of Columbia
                Brazil, State of Minas Gerais
                United States, Kansas
                United States, Virginia
                United States, California
                United States, Missouri
                United States, Illinois
                United States, Texas
                Mexico, Yucatan
                Mexico
                Argentina
                Argentina, Santa Fe
                Argentina, Salta Province
                Argentina, Santa Fe, Las Colonias
                Cuba, Havana
```

Figure 3 Shows the search by repeat interface. Once one or more repeat names or sequences (in FASTA format) is entered along with a maximum edit distance (default zero), the search result window appears with the relevant details for that repeat. The windows are scrollable and can be copied.



C:\Users\Helen\Anaconda2\python.exe

```
Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 3

 What criteria would you like to search by?
1: Search by repeat
2: Search by location
3: Search by strain
0: Return to main menu
Please enter the number of the option you want: 2
```

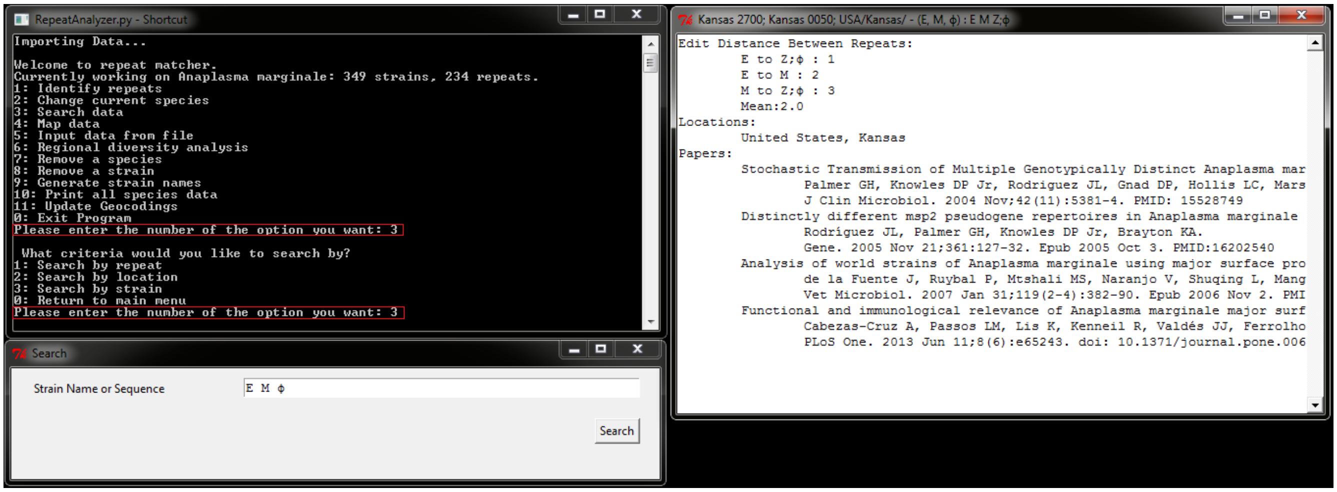
Search

Location: Brazil, Marajó Island

Brazil, Marajó Island

```
Repeats:
        Is9; 78 : tdsssasgqqqessvlsqssqastssqlg
        24 : adsssagnqqqessvlpqsgqastssqls
        25 : adsssasgqqqessvlsqssqastssqlg
        31 : adsssagnqqqessvssqsdastssqlg
        4 : tdsssasgqqqessvlsqsgqastssqlg
        63 : agsssasgqqqessvlsqsgqastssqlg
        27 : adsssasgqqqessvlsqsdqastssqlg
        τ : tdsssasgqqqessvlspsgqastssqlg
        10 : adsssasgqqqessvlspsgqastssqlg
        15 : adsssasgqqqesgvlsqsgqastssqlg
        162 : tdsssasgqqqessvlsqsgqastssllg
Strains:
        Minas-2; Brazil/Marajó Island/E - (78, 24^2, 25, 31) : Is9;78 24 24 25 3
        Brazil/Marajó Island/E - (4, 63, 27) : 4 63 27
        Brazil/Marajó Island/E - ( τ, 10^2, 15) : τ 10 10 15
        Brazil/Marajó Island/E - (162, 63, 27) : 162 63 27
Papers:
        Detection of genetic diversity of Anaplasma marginale isolates in Minas
                Pohl AE, Cabezas-Cruz A, Ribeiro MF, Silveira JA, Silaghi C, Pfi
                Rev Bras Parasitol Vet. 2013 Jan-Mar;22(1):129-35. PMID: 2425295
        Low genetic diversity associated with low prevalence of Anaplasma margin
                Silva JB, Fonseca AH, Barbosa JD, Cabezas-Cruz A, de la Fuente J
                Ticks Tick Borne Dis. 2014 Oct;5(6):801-4. doi: 10.1016/j.ttbdis
```

Figure 4 Shows the Search by location interface. Similar to Figure 3, search by location shows all relevant information for a given location, selected from the alphabetized list in the search window. Locations can be as broad as a country, or as narrow as a county, and broader location will include results from all narrower locations as well. (i.e. the result for Brazil includes results for any Brazilian province or county with data.) If a location does not appear on the list, then there are no genotype reports in that location.



RepeatAnalyzer.py - Shortcut

```
Importing Data...

Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 3

 What criteria would you like to search by?
1: Search by repeat
2: Search by location
3: Search by strain
0: Return to main menu
Please enter the number of the option you want: 3
```

Search

Strain Name or Sequence: E M ϕ

Kansas 2700; Kansas 0050; USA/Kansas/ - (E, M, ϕ) : E M Z;ϕ

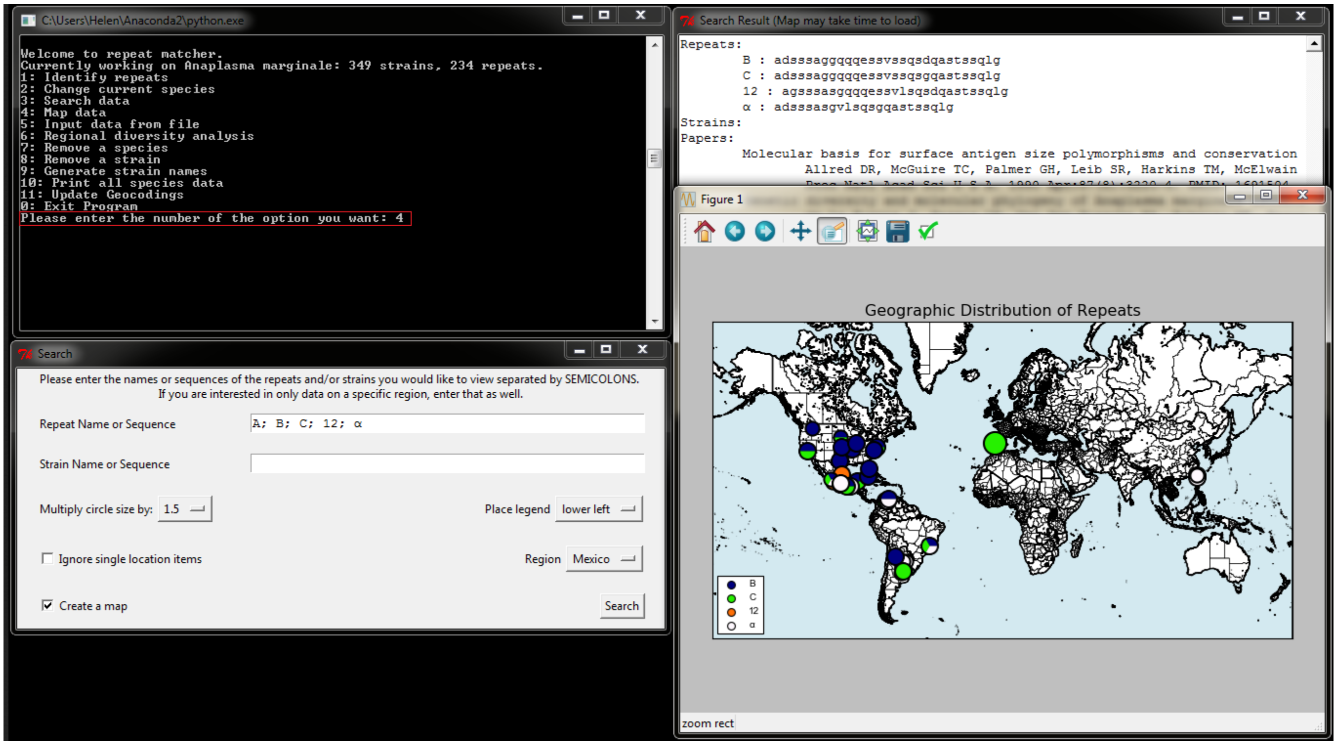
```
Edit Distance Between Repeats:
        E to Z;ϕ : 1
        E to M : 2
        M to Z;ϕ : 3
        Mean:2.0
Locations:
        United States, Kansas
Papers:
        Stochastic Transmission of Multiple Genotypically Distinct Anaplasma mar
                Palmer GH, Knowles DP Jr, Rodriguez JL, Gnad DP, Hollis LC, Mars
                J Clin Microbiol. 2004 Nov;42(11):5381-4. PMID: 15528749
        Distinctly different msp2 pseudogene repertoires in Anaplasma marginale
                Rodríguez JL, Palmer GH, Knowles DP Jr, Brayton KA.
                Gene. 2005 Nov 21;361:127-32. Epub 2005 Oct 3. PMID:16202540
        Analysis of world strains of Anaplasma marginale using major surface pro
                de la Fuente J, Ruybal P, Mtshali MS, Naranjo V, Shuqing L, Mang
                Vet Microbiol. 2007 Jan 31;119(2-4):382-90. Epub 2006 Nov 2. PMI
        Functional and immunological relevance of Anaplasma marginale major surf
                Cabezas-Cruz A, Passos LM, Lis K, Kenneil R, Valdés JJ, Ferrolho
                PLoS One. 2013 Jun 11;8(6):e65243. doi: 10.1371/journal.pone.006
```

Figure 5 Shows the Search by Strain interface. It can accept either a single strain name or sequence, or multiple in FASTA. For each input it returns the edit distances for all repeats in the strain, all locations it has been reported, and the associated publications.

4. Map data

Input: Any combination of: A list of repeats by name or sequence, a list of genotypes by name or sequence, and a location. Lists are separated by semicolons. Repeats in a genotype are separated by whitespace. Lists may be replaced by the word 'All' to include all entries, or left blank to indicate none. There is also a dropdown menu to select where the legend is placed and marker scale.

Output: A printout of data on the mapped entities and a world map showing where those entities occur. The map can be zoomed and it can be saved.



C:\Users\Helen\Anaconda2\python.exe

```
Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 4
```

Search

Please enter the names or sequences of the repeats and/or strains you would like to view separated by SEMICOLONS.
If you are interested in only data on a specific region, enter that as well.

Repeat Name or Sequence: A; B; C; 12; α

Strain Name or Sequence:

Multiply circle size by: 1.5

Place legend: lower left

☐ Ignore single location items

Region: Mexico

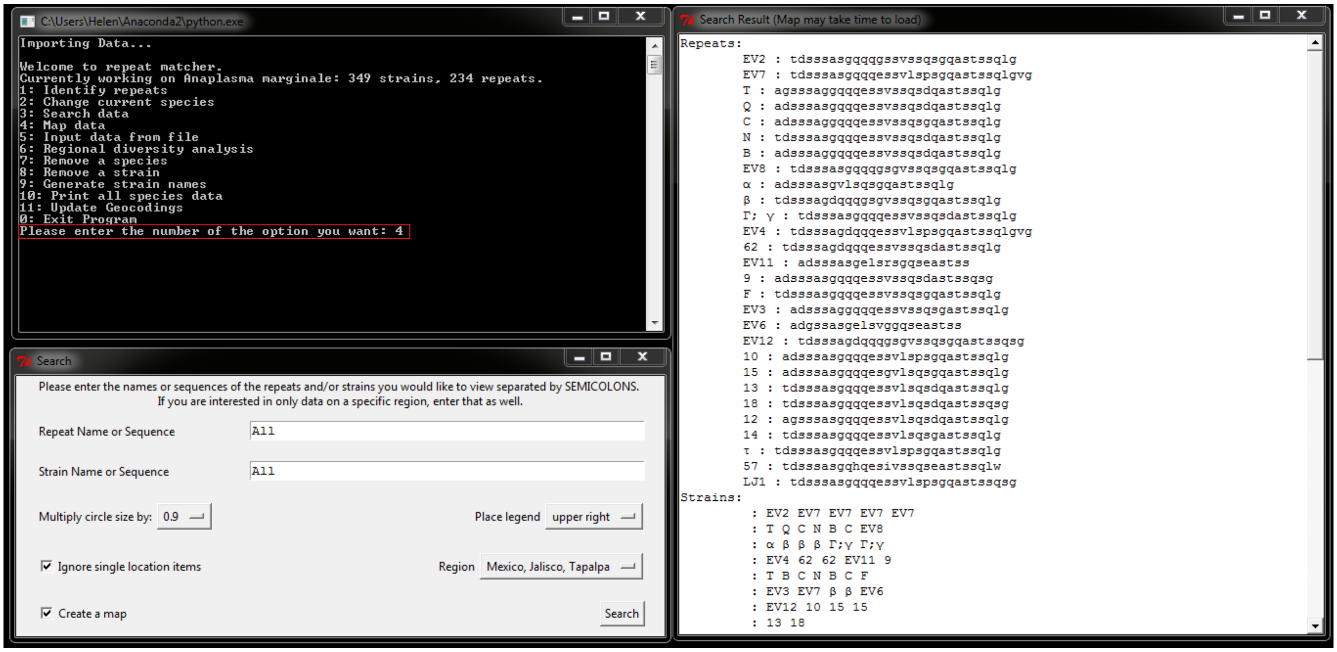
☑ Create a map

Search Result (Map may take time to load)

```
Repeats:
        B : adsssaggqqqessvssqsdqastssqlg
        C : adsssaggqqqessvssqsgqastssqlg
        12 : agsssasgqqqessvlsqsdqastssqlg
        α : adsssasgvlsqsgqastssqlg
Strains:
Papers:
        Molecular basis for surface antigen size polymorphisms and conservation
                Allred DR, McGuire TC, Palmer GH, Leib SR, Harkins TM, McElwain
```

Figure 6 Shows a sample mapping query and result. Each map query can be modified by a number of

options including region (results include only repeats/genotypes present in that region, but all locations they occur in the world), marker scale, legend location and the specific repeats/genotypes of interest. The map plot can be scaled, zoomed, panned and saved to a separate file using the options on the top left of the panel.



```
Importing Data...

Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 4
```

Please enter the names or sequences of the repeats and/or strains you would like to view separated by SEMICOLONS.
If you are interested in only data on a specific region, enter that as well.

Repeat Name or Sequence: All

Strain Name or Sequence: All

Multiply circle size by: 0.9

Place legend: upper right

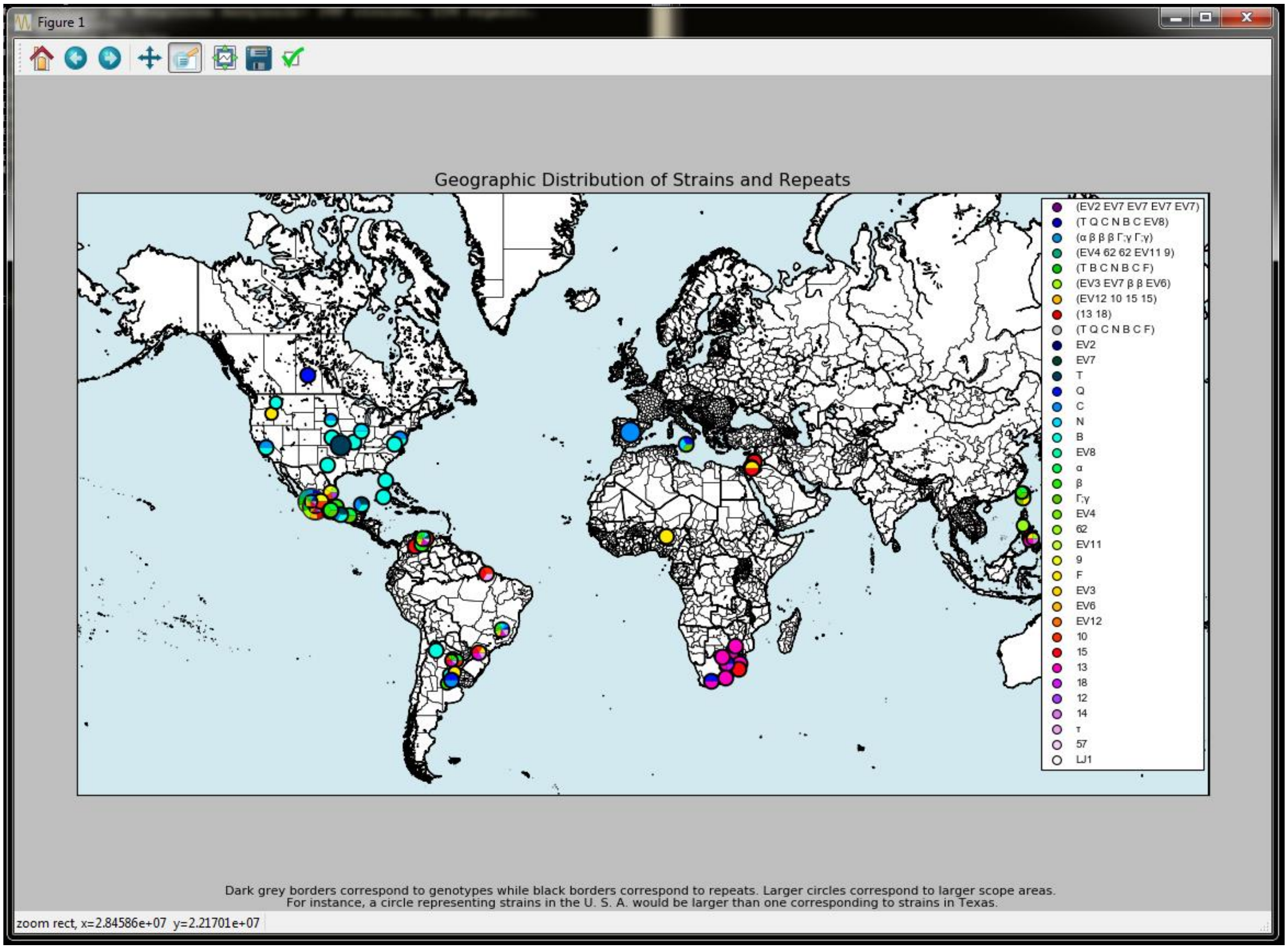
☑ Ignore single location items

Region: Mexico, Jalisco, Tapalpa

☑ Create a map

```
Repeats:
        EV2 : tdsssasgqqqgssvssqsgqastssqlg
        EV7 : tdsssasgqqqessvlspsgqastssqlgvg
        T : agsssaggqqqessvssqsdqastssqlg
        Q : adsssasgqqqessvssqsdqastssqlg
        C : adsssaggqqqessvssqsgqastssqlg
        N : tdsssasgqqqessvssqsdqastssqlg
        B : adsssaggqqqessvssqsdqastssqlg
        EV8 : tdsssasgqqqgsgvssqsgqastssqlg
        α : adsssasgvlsqsgqastssqlg
        β : tdsssagdqqqgsgvssqsgqastssqlg
        Γ; γ : tdsssasgqqqessvssqsdastssqlg
        EV4 : tdsssagdqqqessvlspsgqastssqlgvg
        62 : tdsssagdqqqessvssqsdastssqlg
        EV11 : adsssasgelsrsgqseastss
        9 : adsssasgqqqessvssqsdastssqsg
        F : tdsssasgqqqessvssqsgqastssqlg
        EV3 : adsssaggqqqessvssqsgastssqlg
        EV6 : adgssasgelsvggqseastss
        EV12 : tdsssagdqqqgsgvssqsgqastssqsg
        10 : adsssasgqqqessvlspsgqastssqlg
        15 : adsssasgqqqesgvlsqsgqastssqlg
        13 : tdsssasgqqqessvlsqsdqastssqlg
        18 : tdsssasgqqqessvlsqsdqastssqsg
        12 : agsssasgqqqessvlsqsdqastssqlg
        14 : tdsssasgqqqessvlsqsgastssqlg
        τ : tdsssasgqqqessvlspsgqastssqlg
        57 : tdsssasgqhqesivssqseastssqlw
        LJ1 : tdsssasgqqqessvlspsgqastssqsg
Strains:
         : EV2 EV7 EV7 EV7 EV7
         : T Q C N B C EV8
         : α β β β Γ;γ Γ;γ
         : EV4 62 62 EV11 9
         : T B C N B C F
         : EV3 EV7 β β EV6
         : EV12 10 15 15
         : 13 18
```

Figure 7 Shows a sample mapping query. When sequence is set to 'All' the map will show all results for the specified region. There is also an option to ignore items that appear at only a single location.



Dark grey borders correspond to genotypes while black borders correspond to repeats. Larger circles correspond to larger scope areas.
For instance, a circle representing strains in the U. S. A. would be larger than one corresponding to strains in Texas.

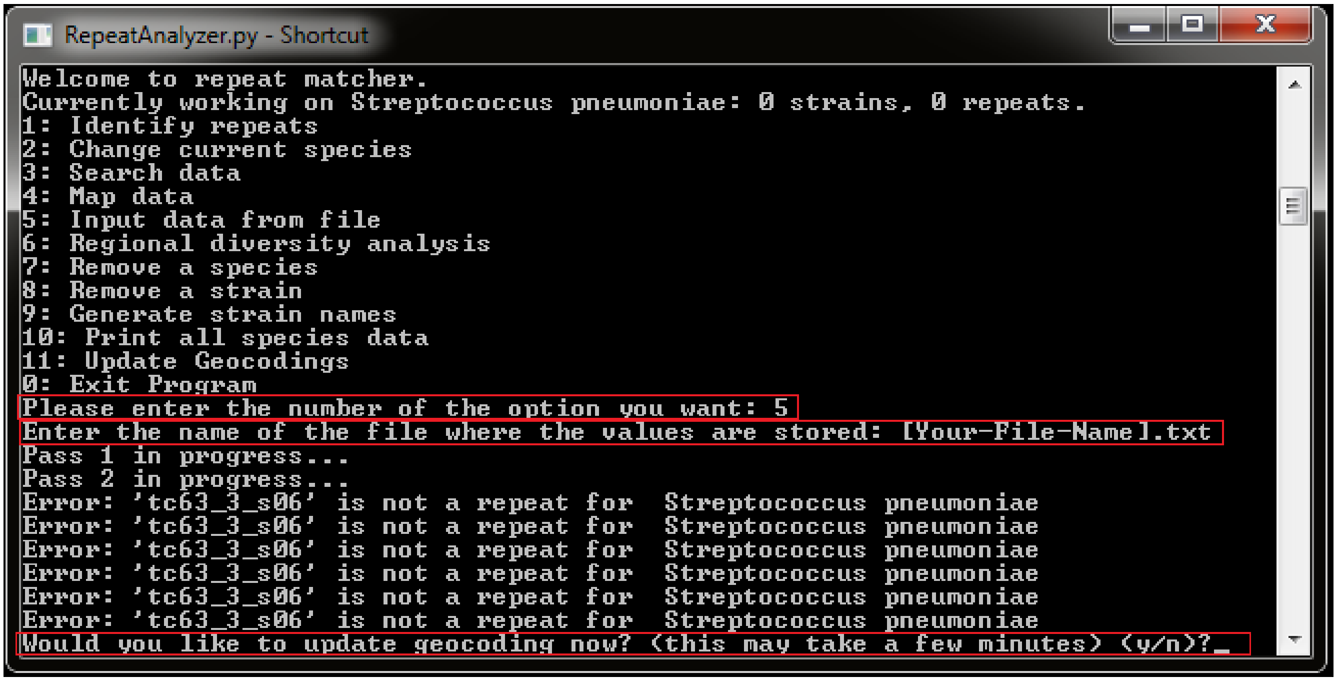
```
zoom rect, x=2.84586e+07  y=2.21701e+07
```

Figure 8 Shows a sample mapping result for the query in figure 7.

5. Input data from file

Input: The name of a file formatted as described in the Sample Input File Format section.

Output: Any error in the input file will be noted. The main menu header will change as appropriate.



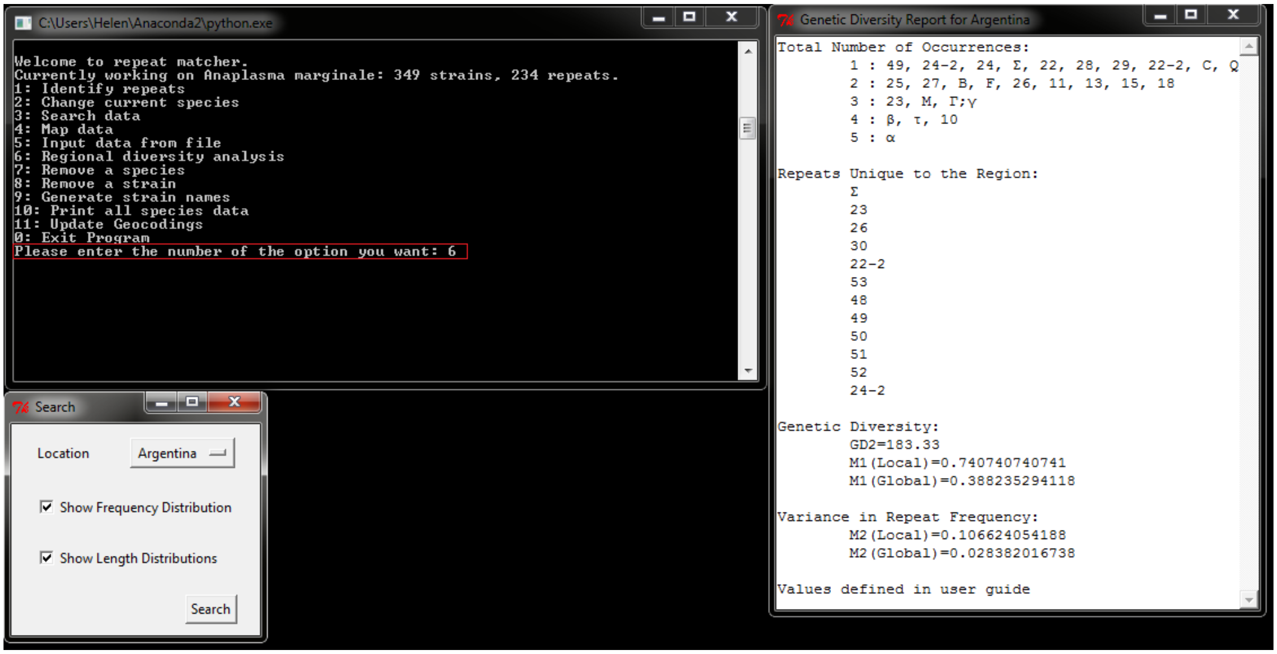
```
Welcome to repeat matcher.
Currently working on Streptococcus pneumoniae: 0 strains, 0 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 5
Enter the name of the file where the values are stored: [Your-File-Name].txt
Pass 1 in progress...
Pass 2 in progress...
Error: 'tc63_3_s06' is not a repeat for  Streptococcus pneumoniae
Error: 'tc63_3_s06' is not a repeat for  Streptococcus pneumoniae
Error: 'tc63_3_s06' is not a repeat for  Streptococcus pneumoniae
Error: 'tc63_3_s06' is not a repeat for  Streptococcus pneumoniae
Error: 'tc63_3_s06' is not a repeat for  Streptococcus pneumoniae
Error: 'tc63_3_s06' is not a repeat for  Streptococcus pneumoniae
Would you like to update geocoding now? (this may take a few minutes) (y/n)?_
```

Figure 9 Shows the sequence of commands to input new data for a species. The first command opens the data input menu. While the second allows the user to enter the name of the file where the data is stored. Note that if there are any errors which prevent the file from reading, the line number of the error will be shown. If any repeats are listed in strains, but never specified in the repeats section, the error shown above the third command will print, and that strain will be omitted, but the rest of the file will be read in correctly. Finally, the third command may either be entered 'y' if a stable internet connection is available, or anything else, if not. This step is required for certain program features to work correctly, and can be done later if it is skipped now (function 11).

6. Regional diversity analysis

Input: Region, selected from a dropdown menu and checkboxes for additional plots at needed.

Output: A printout with all diversity scores, repeat frequencies, unique repeats and (if selected) appropriate plots.


C:\Users\Helen\Anaconda2\python.exe
Search
Genetic Diversity Report for Argentina
Search

```
Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 6
```

Location Argentina

☑ Show Frequency Distribution

☑ Show Length Distributions

```
Total Number of Occurrences:
        1 : 49, 24-2, 24, Σ, 22, 28, 29, 22-2, C, Q
        2 : 25, 27, B, F, 26, 11, 13, 15, 18
        3 : 23, M, Γ;γ
        4 : β, τ, 10
        5 : α

Repeats Unique to the Region:
        Σ
        23
        26
        30
        22-2
        53
        48
        49
        50
        51
        52
        24-2

Genetic Diversity:
        GD2=183.33
        M1(Local)=0.740740740741
        M1(Global)=0.388235294118

Variance in Repeat Frequency:
        M2(Local)=0.106624054188
        M2(Global)=0.028382016738

Values defined in user guide
```

Figure 10 Shows the regional diversity analysis interface and a sample query. For regional diversity analysis, you can select the region of interest as well as whether or not to produce various plots (shown in Figure 11). The result has several sections including, how many genotypes each repeat in the region occurs in, a list of all repeats that occur nowhere outside the region, and a list of the various diversity scores as defined in Table 1.

Metric Name	***Formula***
GD2	$\frac{Total\ \#\ Unique\ SSRs}{\#\ Genotypes}$
GDM1 (Local)	$Avg\left(\frac{\#\ Unique\ SSRs\ in\ Genotype\ i}{Length(Genotype\ i)}\right)$
GDM1 (Global)	$\frac{Total\ \#\ Unique\ SSRs}{Total\ Length(All\ Genotypes)}$
GDM2 (Local)	$Avg\left(Deviation\left(\frac{Frequency(SSR\ i\ in\ Genotype\ j)}{Length(Genotype\ j)}\right)\right)$
GDM2 (Global)	$Deviation\left(\frac{Frequency(SSR\ i)}{Total\ Length(All\ Genotypes)}\right)$

Table 1. Metrics used to calculate genetic diversity of a geographic region.

Length(Genotype) = the number of SSRs in in that genotype

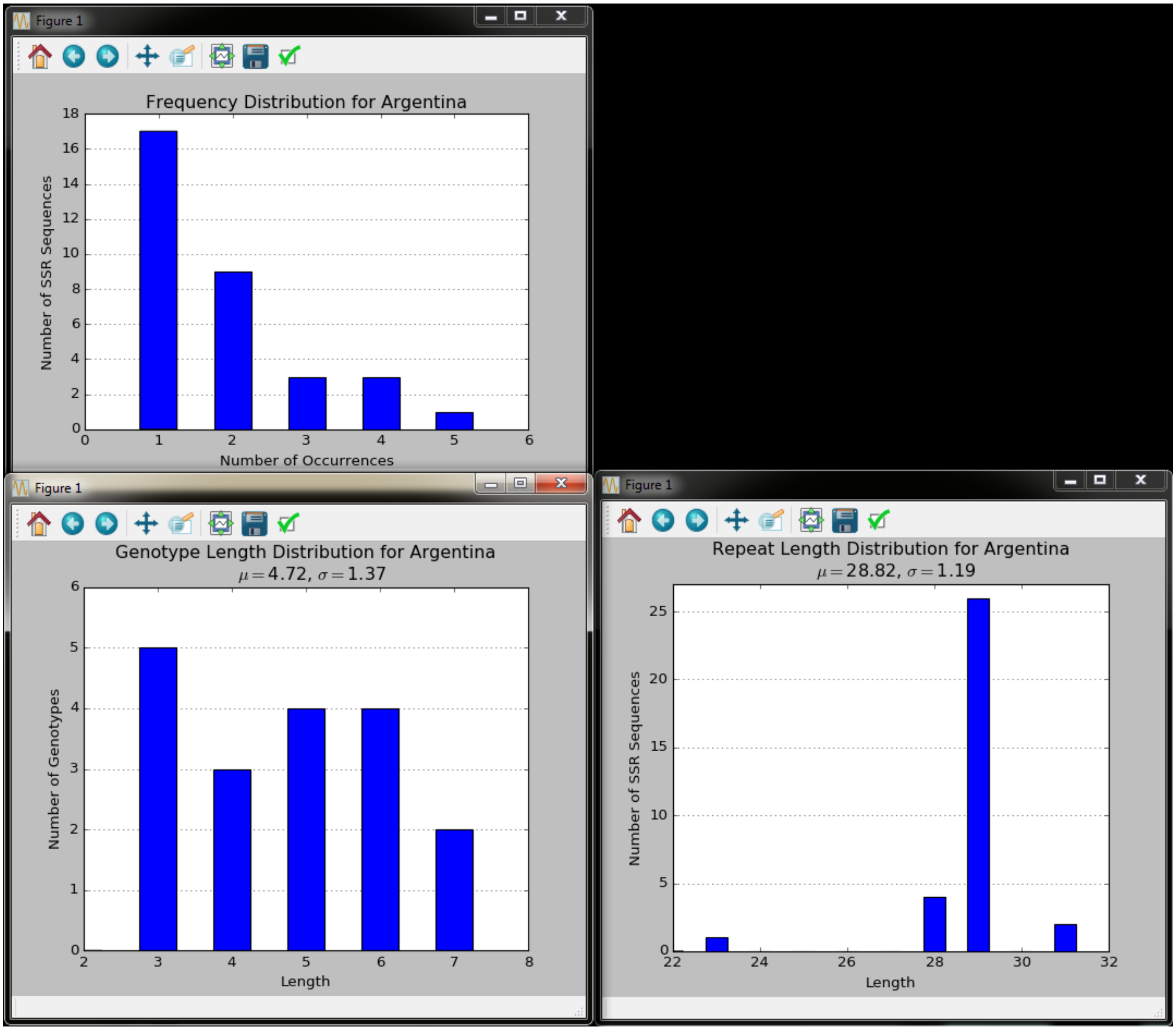
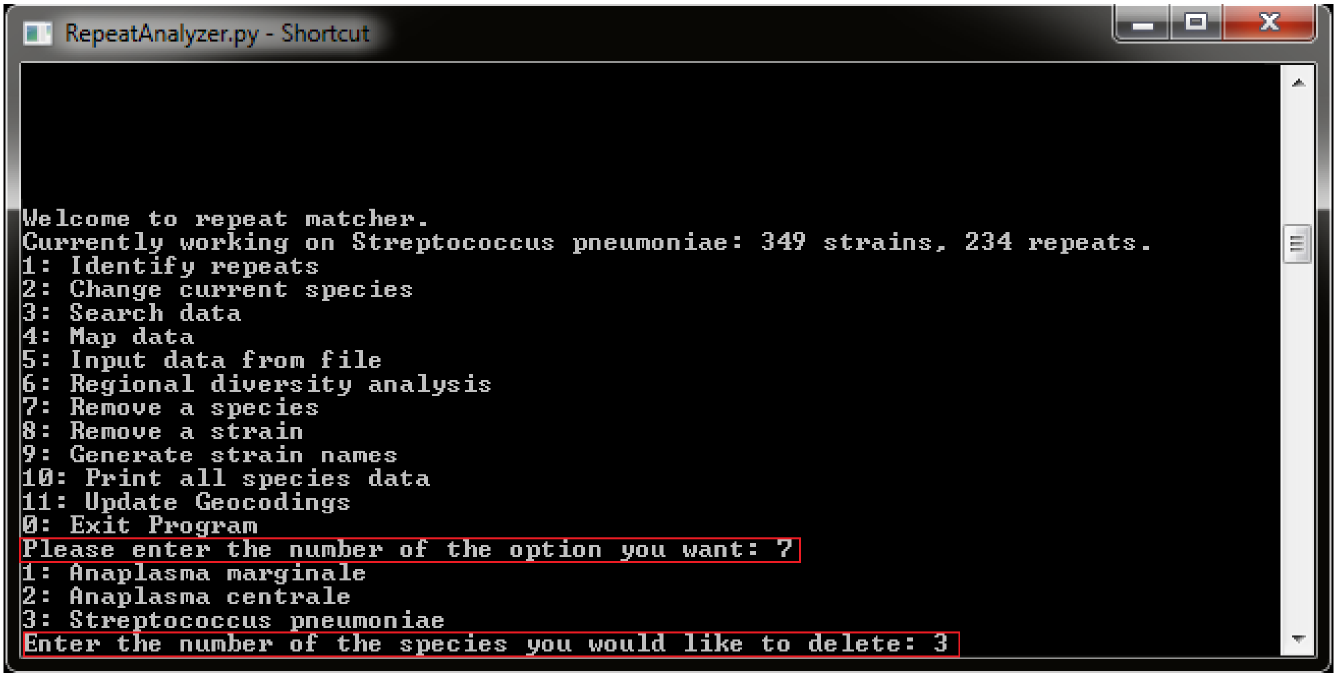


Figure 11 Shows the plots produced by the regional diversity analysis in Figure 10.

7. Remove a species

Input: The ID of the species to be removed (as shown on-screen)

Output: None



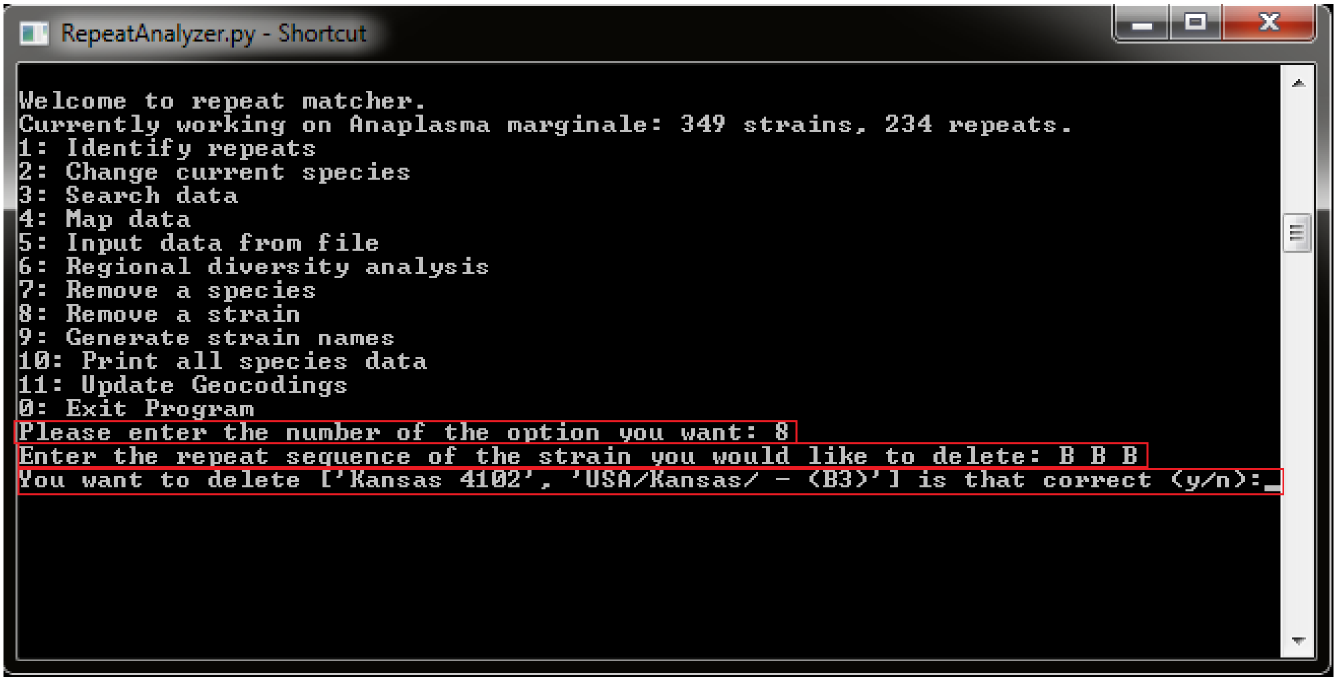
```
Welcome to repeat matcher.
Currently working on Streptococcus pneumoniae: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 7
1: Anaplasma marginale
2: Anaplasma centrale
3: Streptococcus pneumoniae
Enter the number of the species you would like to delete: 3
```

Figure 12 Shows the series of commands to remove a species. The first command opens the species removal menu, while the second selects the number of the species to remove.

8. Remove a strain

Input: The repeat sequence of the strain to remove.

Output: The main menu header will change as appropriate.



```
Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 8
Enter the repeat sequence of the strain you would like to delete: B B B
You want to delete ['Kansas 4102', 'USA/Kansas/ - (B3)'] is that correct (y/n):_
```

Figure 13 Shows the series of commands to remove a strain. The first command opens the strain removal menu. The second command allows the user to enter the repeat sequence of the strain to be removed. Finally, the third command may either be entered 'y' if the correct strain was selected.

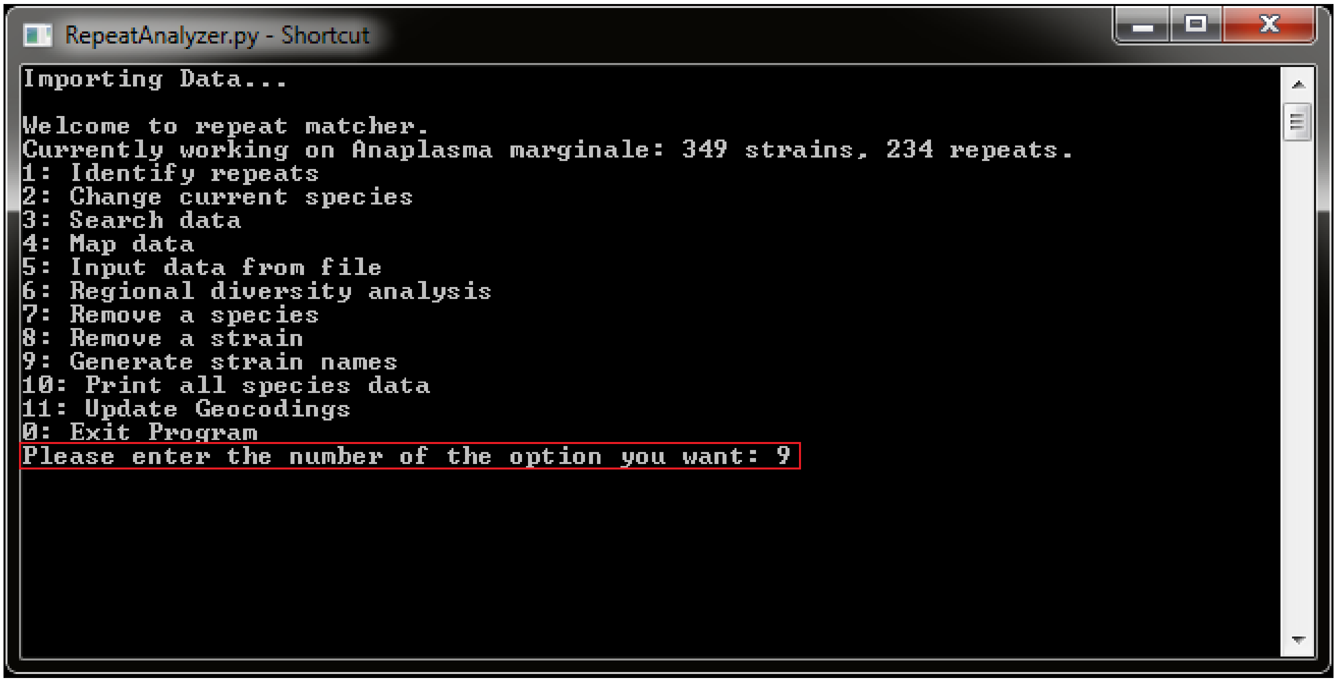
9. Generate strain names

Input: None

Output: None

Note: This function will generate any names for newly input data where the genotype name was listed as year, animalID. There is no direct input or output.

While typically this step is done when menu option 5 is run, it can only be completed after geocoding (which requires an internet connection) and so, it may optionally be skipped and run later, via this command.



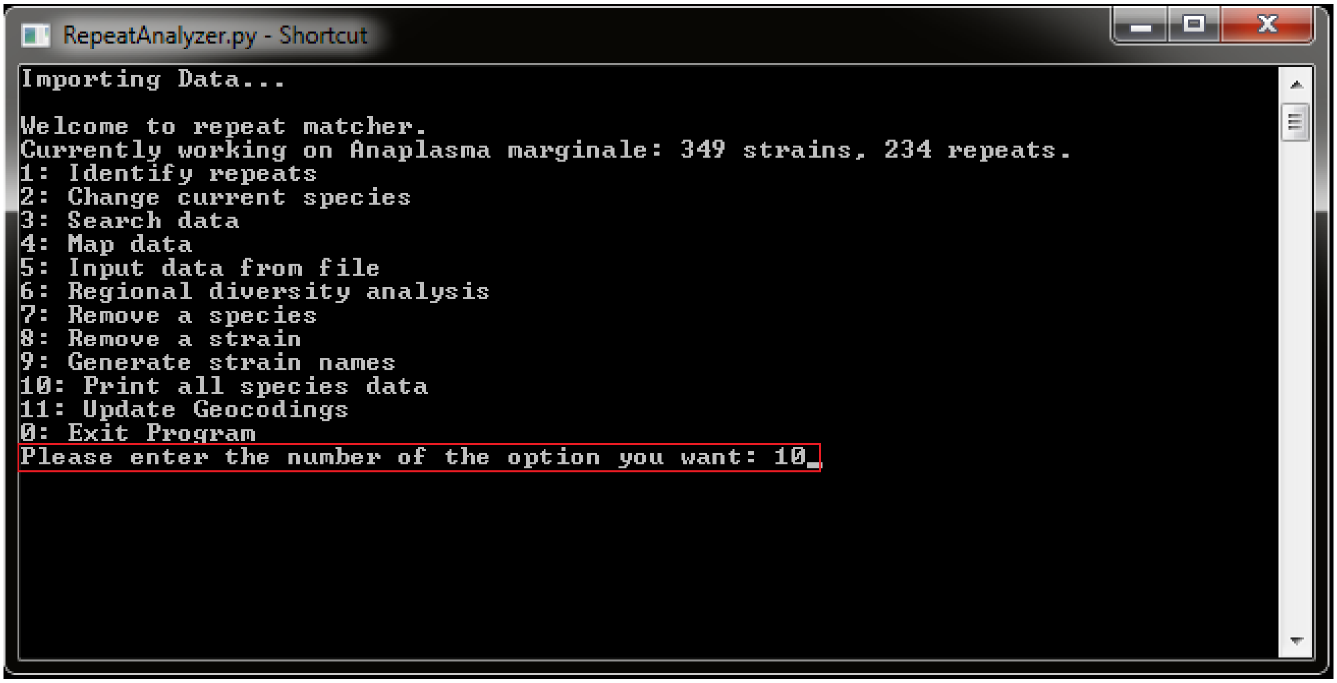
```
Importing Data...

Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 9
```

Figure 14 Shows the command to automatically generate species name for species input with the correct information.

10. Print species data

Input: None

Output: A text printout with a summary of all data, and a spreadsheet with all repeats and their sequences for easy manipulation. These will be located in the same folder as RepeatAnalyzer.py.



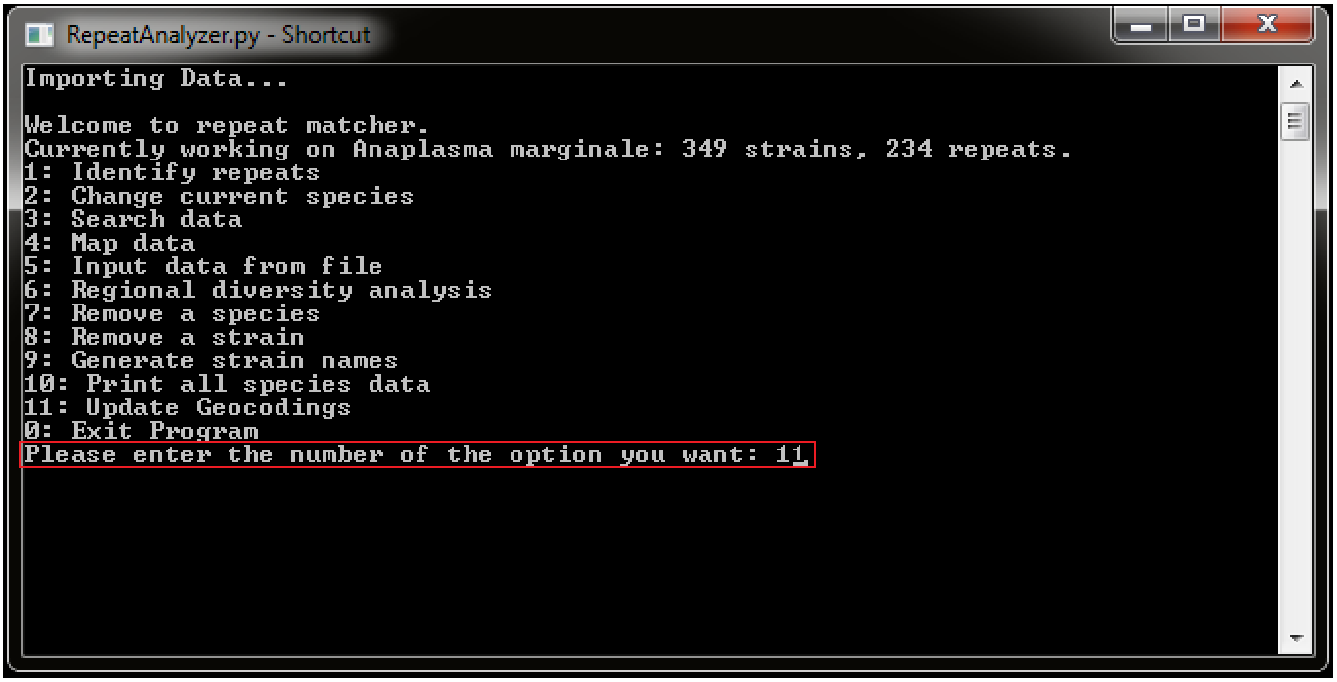
```
Importing Data...

Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 10_
```

Figure 15 Shows the command to print a summary of all species data to an external file.

11. Update geocodings

Input: None

Output: Formats the names of all locations to be uniform (so that USA, U. S. A., and United States all point to the same place) and gets the coordinates for locations that have none. This must be done at least once after adding data for mapping to function properly. Note: there is a y/n option to do this after adding new data as well. An internet connection is required.



```
Importing Data...

Welcome to repeat matcher.
Currently working on Anaplasma marginale: 349 strains, 234 repeats.
1: Identify repeats
2: Change current species
3: Search data
4: Map data
5: Input data from file
6: Regional diversity analysis
7: Remove a species
8: Remove a strain
9: Generate strain names
10: Print all species data
11: Update Geocodings
0: Exit Program
Please enter the number of the option you want: 11
```

Figure 16 Shows the command to update geographic codings.

Troubleshooting

If you have experienced problems running RepeatAnalyzer there is likely a problem with your installation, or a conflict with another program (such as an older version of Python) installed on your system. We can help resolve certain issues, but this will be considerably faster with the error messages specific to your setup. If possible, send any troubleshooting requests along with the error log acquired in the following way.

To give us more information about what your specific problem is, open a program on your computer called “spyder” (part of the Anaconda environment) and in that program open the RepeatAnalyzer.py file. Hit the 'f5' key and confirm the dialog that appears. An error message should appear on the lower right-hand side of the screen. Copy all of the text there (it may be a lot, so be sure to scroll up if needed) and send it to us.

Email information about configuration problems/conflicts to helen.catanese@wsu.edu and we will attempt to resolve the issue.