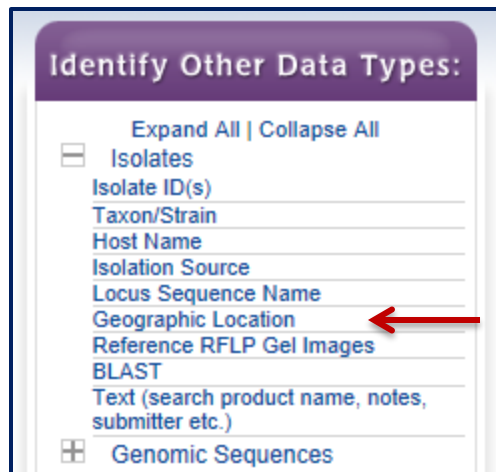
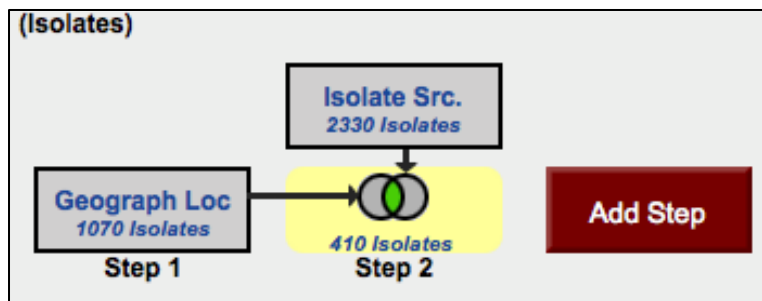


Exploring Isolate Data

1. Exploring isolates in *Cryptosporidium* and using the alignment tool. (<http://www.cryptodb.org>)
 - a. Identify all *Cryptosporidium* isolates from Europe. (hint: search for isolates by geographic location in the “Identify Other Data Types” section).



- b. How many of the *Cryptosporidium* isolates collected in Europe were isolated from feces? (hint: add another isolate search step - isolation source).



- c. What is the general distribution of these isolates in Europe? (hint: you can do this quickly in two ways: sort the geographic location column by clicking on the sort arrows, then look at the represented countries; or use the “Isolate Geographical Location” tab to view a map and results summary table).

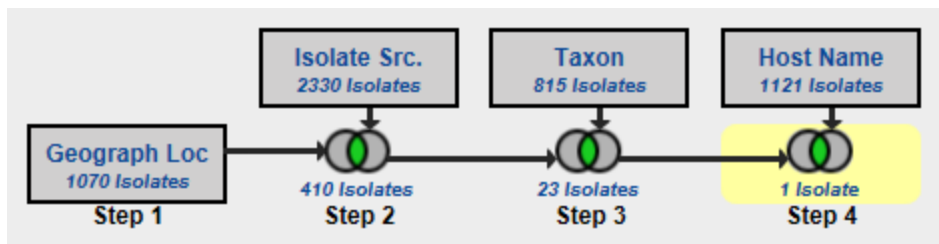
Sort by clicking on the arrows

Isolate #	Geographic Location	Organism	Strain/Isolate Name	Host	Isolation Date
AB242224	Serbia	Cryptosporidium parvum	#6	Unknown	fecal sample
AB242225	Serbia	Cryptosporidium parvum	#24	Unknown	fecal sample
AB242226	Serbia	Cryptosporidium parvum	#42	Unknown	fecal sample
AB242227	Serbia				

Country	Number of Isolates	Isolate Type	Latitude	Longitude
Belgium	1	Sequencing Typed	50.503887	4.469936
Czech Republic	73	Sequencing Typed	49.817492	15.472962
Germany	90	Sequencing Typed	51.165691	10.451526
Ireland	4	Sequencing Typed	53.41291	-8.24389
Italy	11	Sequencing Typed	41.87194	12.56738
Lithuania	1	Sequencing Typed	55.169438	23.881275
Netherlands	41	Sequencing Typed	52.132633	5.291266

- d. Out of those in step ‘b’, how many are unclassified *Cryptosporidium* species? (hint: add another isolate search step and select taxon/strain then select the unclassified isolates)

- e. How many of step ‘d’ isolates originated from humans?



- f. How many of the isolates in step 'b' were typed using GP40/15 (GP60)? (hint: you can insert a step within a strategy. Click "edit" on the step of interest then select "Insert step before").

The screenshot shows a software interface for managing strategies. On the left, a strategy diagram shows 'Geograph Loc' (Step 1) leading to 'Isolate Src.' (Step 2), which leads to 'STEP 3: Locus Sequence Name'. The 'Insert Step Before' button is highlighted with a red box. The detailed view of Step 3 shows the following information:

- STEP 3 : Locus Sequence Name**
- Locus Sequence Name : sporozoite antigen gp40/15 (gp60)
- Locus Sequence wildcard search : N/A
- Results: 1476 Isolates
- Give this search a weight

Below the strategy view, a table shows the first 8 isolates from Step 3:

Isolate Id	Organism	Strain/Isolate Name	Host	Geographic Location
AB242224	Cryptosporidium parvum	#6	Unknown	Serbia
AB242225	Cryptosporidium parvum	#24	Unknown	Serbia

- g. Compare the first 8 isolates using the multiple sequence alignment tool (ClustalW). Do you see any sequences with insertions or deletions?

The screenshot shows the ClustalW interface with a list of 8 isolates from Step 3:

Isolate Id	Organism	Strain/Isolate Name	Host	Geographic Location
AB242224	Cryptosporidium parvum	#6	Unknown	Serbia
AB242225	Cryptosporidium parvum	#24	Unknown	Serbia
AB242226	Cryptosporidium parvum	#42	Unknown	Serbia
AB242227	Cryptosporidium parvum	#58	Unknown	Serbia
AB242228	Cryptosporidium parvum	#80	Unknown	Serbia

Below the table, there is a note: "Please select at least two isolates to run ClustalW. Note: only isolates from a single page (Increase the page size in 'Advanced Paging' to increase the number that can be displayed)." and a button labeled "Run Clustalw on Checked Strains" which is highlighted with a red box.

- h. Create a guide tree based on this alignment. Below the alignment in the output file, are the contents of a .dnd file. The Newick Viewer (link below) uses the dnd file to create a guide tree. Cut and paste the .dnd file beginning with the first open parenthesis into the Newick Viewer string box. Then click View Tree.

<http://www.trex.uqam.ca/index.php?action=newick>

Newick Viewer

Newick Viewer allows you to visualize a tree coded by its Newick string. Hierarchical, Axial and Radial types of tree drawing are available.

Paste your **Newick** string into the window :

```
(
(
AB242224:0.00066,
AB242228:-0.00066)
:0.00126,
(
(
AB242225:0.00000,
AB242227:0.00000)
:0.01298,
(
AB242226:0.08807,
(
AY508960:0.14832,
AY508961:0.13168)
:0.69623)
:0.07248)
:0.02254,
AB242229:-0.00126);
```

Sequences file Pasted

Newick string without brackets (for the sequence file)

.dnd file

```
(
(
AB242224:0.00066,
AB242228:-0.00066)
:0.00126,
(
(
AB242225:0.00000,
AB242227:0.00000)
:0.01298,
(
AB242226:0.08807,
(
AY508960:0.14832,
AY508961:0.13168)
:0.69623)
:0.07248)
:0.02254,
AB242229:-0.00126);
```

- i. Change the isolates that you selected for alignment - how does the tree change? Do isolates from the same country cluster together?

2. Typing an unclassified *Cryptosporidium* isolate. (<http://www.cryptodb.org>)

- a. You have just finished sequencing part of the 18S small subunit ribosomal RNA gene from isolates you retrieved from a *Cryptosporidium* outbreak at a public swimming pool in Uppsala. The sequence was identical from all the isolates and is pasted below. Can you use CryptoDB to get an idea of which reference isolate this is most similar to? (hint: go to the BLAST page in CryptoDB and blast your sequence against the reference isolates).

```
AAGCTCGTAGTTGGATTTCTGTTAATAATTTATATAAAATATTTTGATGAATATTTATAT
AATATTAACATAATTCATATTACTATATATTTTAGTATATGAAATTTTACTTTGAGAAAA
TTAGAGTGCTTAAAGCAGGCATATGCCTTGAATACTCCAGCATGGAATAATATTAAGAT
TTTTATCTTTCTTATTGGTTCTAAGATAAGAATAATGATTAATAGGGACAGTTGGGGGCA
TTTGATTTAACAGTCAGAGGTGAAATCCTTAGATTTGTTAAAGACAACTAATGCGAAA
GCATTTGCCAAGGATGTTTTTCATTAATCAAGAACGAAAGTTAGGGGATCGAAGACGATCA
GATACCGTCGTAGTCTTAACCATAAACTATGCCAACTAGAGATTGGAGGTTGTTCCCTTAC
TCCTTCAGCACCTTA
```

- b. You can get to the BLAST page from the home page (BLAST link under the tool section) or from the isolate searches and select “BLAST”. Configure the BLAST search page: select isolates and make sure only the reference isolates are selected in the target organism window.
- c. Paste the DNA sequence in the input window and select the blastn program. Click on “Get Answer”.

The screenshot shows the CryptoDB BLAST search interface. Red arrows point to the following fields:

- Target Data Type:** Radio buttons for Transcripts, Proteins, Genome, EST, ORF, and **Isolates** (selected).
- BLAST Program:** Radio buttons for **blastn** (selected), blastp, blastx, tblastn, and tblastx.
- Target Organism:** A list of taxonomic groups with checkboxes. **Cryptosporidiidae SSU_18srRNA Reference Isolates** is checked. Other options include Chromerida, Cryptosporidiidae, and Gregarinidae.
- Input Sequence:** A text area containing the DNA sequence from the previous block. A note below states: "Note: only one input sequence allowed. maximum allowed sequence length is 31K bases."
- Expectation value:** Input field with "10".
- Maximum descriptions/alignments (V=B):** Input field with "50".
- Low complexity filter:** Dropdown menu set to "no".
- Advanced Parameters:** A collapsed section.
- Get Answer:** A button to execute the search.
- Give this search a name:** An input field for naming the search.

- d. Explore your results. Based on the similarity which reference isolate is this one closest to?

```
Query= MySeq1
Length=435
Score      E
GQ983352.1| organism=Cryptosporidium hominis | description=Cryp... 785 0.0

> GQ983352.1| organism=Cryptosporidium hominis | description=Cryptosporidium
hominis isolate W15271 small subunit ribosomal
RNA gene, partial sequence
Length=92932

Score = 785 bits (870), Expect = 0.0
Identities = 435/435 (100%), Gaps = 0/435 (0%)
Strand=Plus/Plus

Query 1      AAGCTCGTAGTTGGATTTCTGttaataatttatataaaatattttgatgaatatttatat 60
          |||
Sbjct 1501   AAGCTCGTAGTTGGATTTCTGTTAATAATTTATATAAAATATTTTGATGAATATTTATAT 1560

Query 61     aatattaacataattcatattactatataatatttagtatatGAAATTTACTTTGAGAAAA 120
          |||
Sbjct 1561   AATATTAACATAATTCATATTACTATATATTTTAGTATATGAAATTTACTTTGAGAAAA 1620

Query 121    TTAGAGTGCTTAAAGCAGGCATATGCCTTGAATACTCCAGCATGGAATAATATTAAAGAT 180
          |||
Sbjct 1621   TTAGAGTGCTTAAAGCAGGCATATGCCTTGAATACTCCAGCATGGAATAATATTAAAGAT 1680

Query 181    TTTTATCTTTCTTATTGGTTCTAAGATAAGAATAATGATTAATAGGGACAGTTGGGGGCA 240
          |||
Sbjct 1681   TTTTATCTTTCTTATTGGTTCTAAGATAAGAATAATGATTAATAGGGACAGTTGGGGGCA 1740
```