# # Exercise 1

Name: Aris Podotas

University: National and Kapodistrain University of Athens

Program: Data Science and Information Technologies Master’s Program

Specialty: Bioinformatics and Biomedical data

Lesson: Bioinformatics

File type: Markdown (converted to pdf)

Table of Contents

[# Exercise 1 1](#_Toc179637402)

[# Requirements 2](#_Toc179637403)

[# Analysis 4](#_Toc179637404)

[## Outbreak 14 4](#_Toc179637405)

[### *Blastn* 4](#_Toc179637406)

[### *Megablast* 6](#_Toc179637407)

[## Outbreak 15 7](#_Toc179637408)

[### *Blastn* 7](#_Toc179637409)

[### *Megablast* 8](#_Toc179637410)

[# Conclusions 9](#_Toc179637411)

[## Differences in the algorithms 9](#_Toc179637412)

[## Outbreak 14 10](#_Toc179637413)

[## Outbreak 15 13](#_Toc179637414)

[## Final statement 14](#_Toc179637415)

[# Citations 14](#_Toc179637416)

# # Requirements

The requirements have been slightly rephrased for brevity.

In a hypothetical scenario many people in a city suddenly come down with a serious illness. All the victims have in

common is that they were all in a downtown pedestrian mall at a certain time five days before. Could terrorists have

released a cloud of viruses or bacteria from a vehicle downwind of the mall? You work for the Centers for Disease

Control and Prevention, and you have to find out.

A sample of non-human DNA (bacterial or viral) has been isolated from the victims. Identify the DNA sample as

well as you can. Some of the DNA molecules are very short, and have been partially degraded. You will notice that

the sequence is sprinkled with Ns, “N” stands for “nucleotide” and means that the nucleotide at that position could

not be determined.

Some judgment is called for as you interpret your results. First, everyone has bacteria and viruses in his or her body,

and sometimes they can cause disease. However, we are looking for exotic pathogens with bioterrorism potential

(e.g., anthrax or smallpox rather than the common cold). Even AIDS, although it is deadly, would not work as a

bioterror weapon because the disease develops too slowly and the virus is too hard to disseminate. For the purposes

of this exercise, we will not consider a pathogen a

bioterror agent unless it is listed as a potential agent on the Centers for Disease Control and Prevention Web site at

https://emergency.cdc.gov/agent/agentlist.asp .

Second, organisms that are evolutionarily related have similar DNA, which might lead you to sound a false alarm.

For example, say you find the following when you do a BLAST search on a certain DNA sample:

Bacillus subtilis is a harmless and very common soil bacterium. It is closely related to Bacillus anthracis. Bacillus

anthracis causes anthrax, and is a dangerous bioterror weapon. Note from the similarity score (second column from

the right) that Bacillus subtilis DNA is far more similar to the sample than Bacillus anthracis DNA is. Unless one of

your samples gives a stronger indication of Bacillus anthracis than this, the mention of B. anthracis in the output is

probably just due to genetic similarities between it and B. subtilis.

1. Analyze the samples

>outbreak14

GCCGAGTTAGTCTTGTGCTNACGGAACTTATTGTATGAGTANTGATTTGAAAGAGCTANANTTAAAA

AATCACTAATNAATNTAAGAGCGGACTTAACNAGCGTAAAACTGTCTTACTAATTAATTGTCAGTTA

GCTCGTTCAGGTAATGGTTCCTANCGGNCAATGCAGGAAGAGTTCTACCTGGAACTGANAGACCGC

TGGCGGTGACAACACACTACGTCAAAATAAGA

>outbreak15

TAGTCTTGTGCTNACGGAACTTATTTATGAGGTACCCACCGANTCTGAAAACCGCTAATANAGCACT

TTAAAAATAAGAGCAGAATGGGATTTAAGGATAG

separately using both *megablast* and *blastn* and to determine if there is any evidence of bioterror agents.

2. Check the CDC Web site at https://emergency.cdc.gov/agent/agentlist.asp .

to see if the CDC considers any found organism to be a potential weapon. If you’ve found a bioterror agent, research

it on the CDC site so you can describe its effects on humans.

3. The health effects of many pathogenic bacteria are briefly described on the NCBI Genomes Web site at

<http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>. Click on a species name to see its information. It also might be

helpful to do a general Google search.

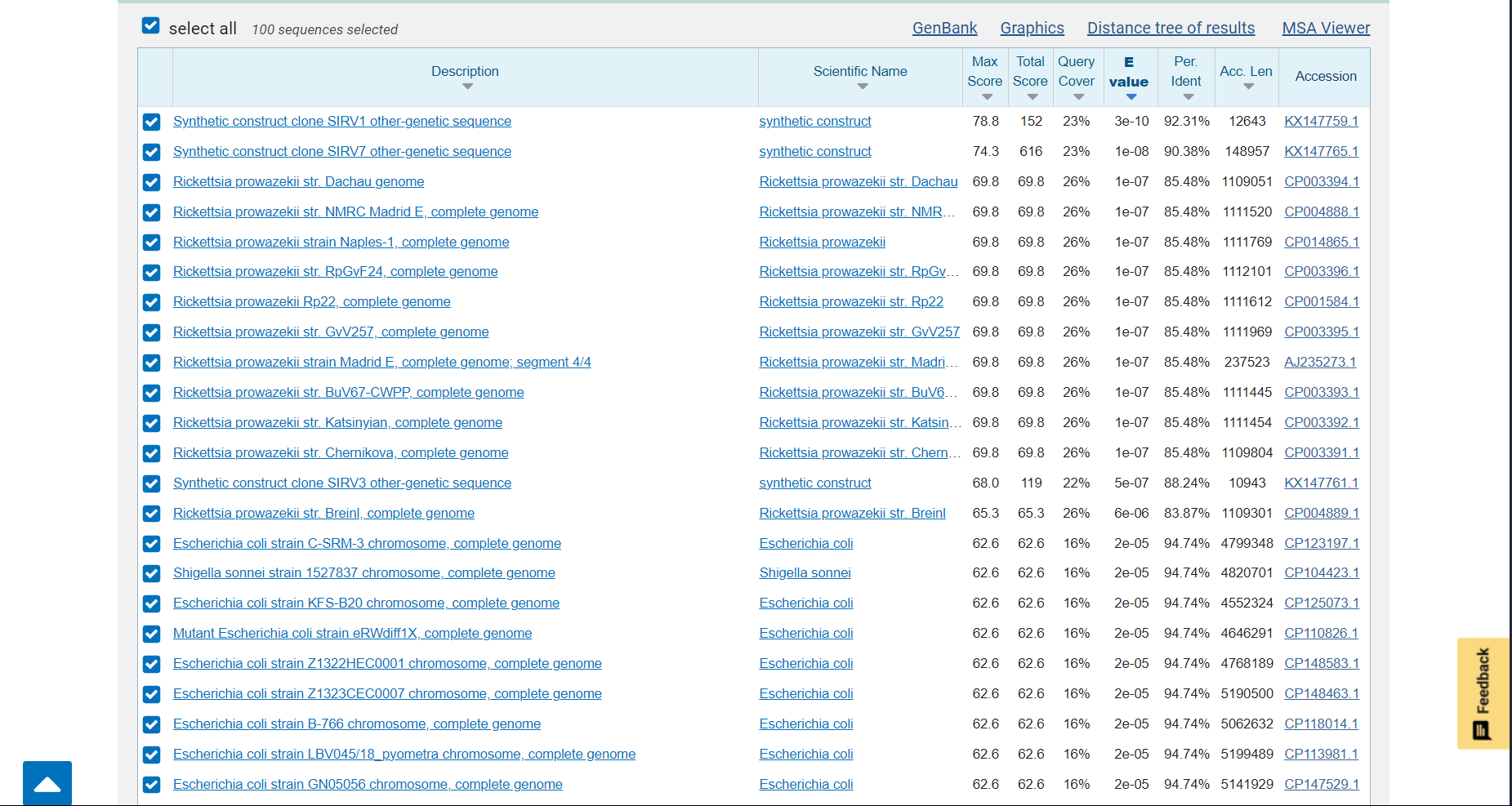
SEND SOLUTIONS (for M.Reczko exercises) ONLY TO:

mareczko@di.uoa.gr

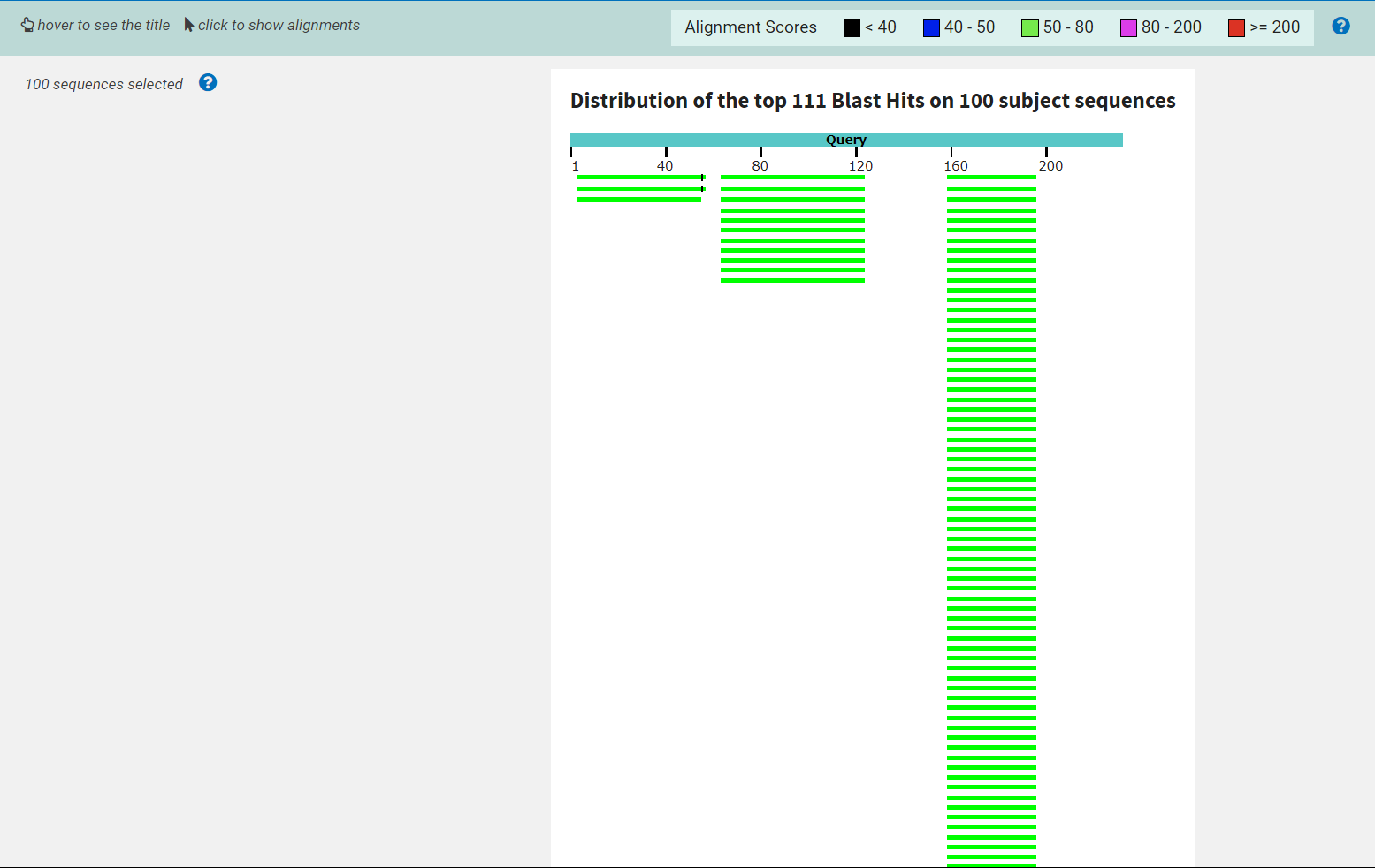
# # Analysis

## ## Outbreak 14

### ### *Blastn*



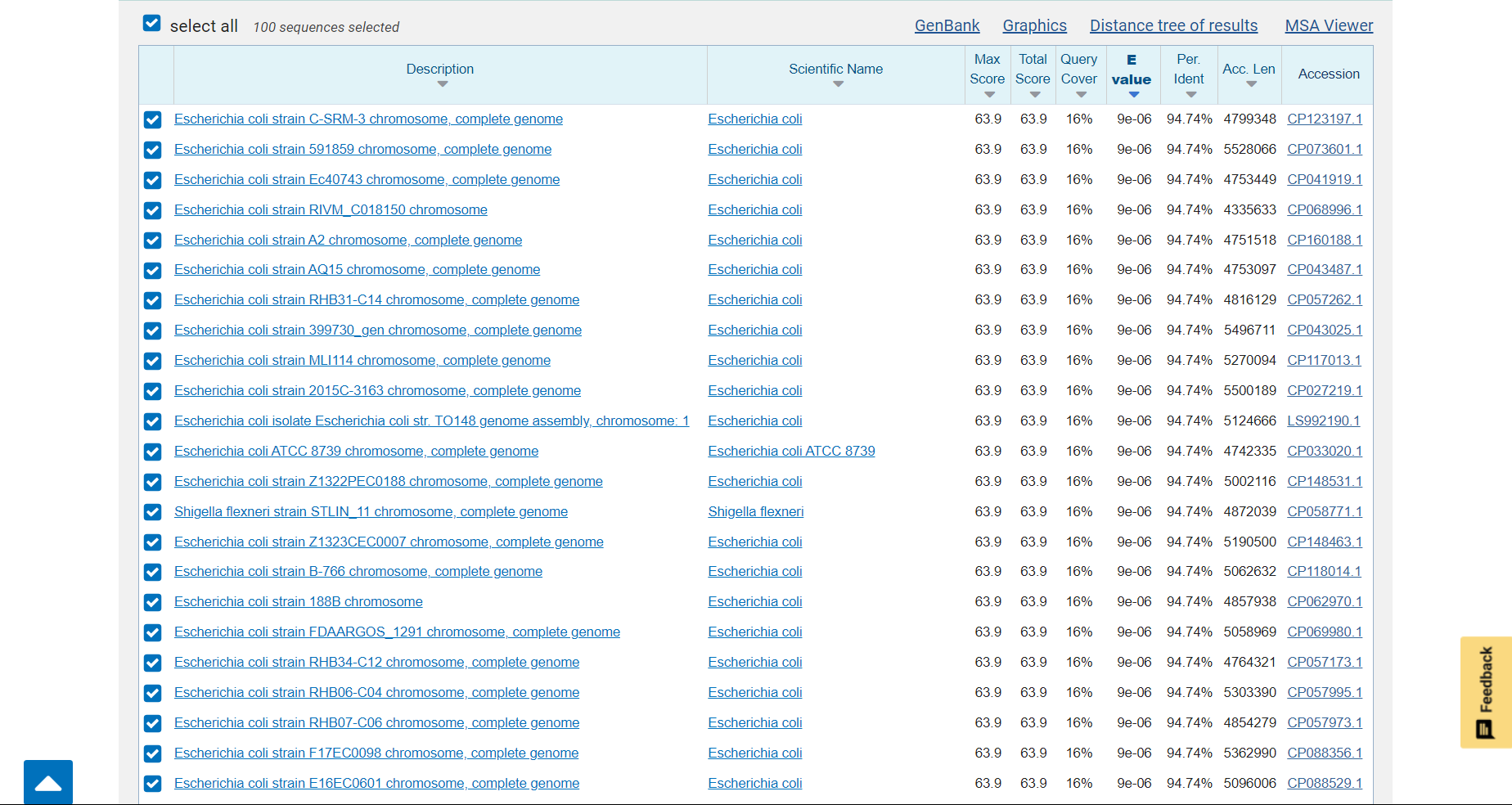
***Figure 1a:*** *Image of the results from blastn for Outbreak 14.*



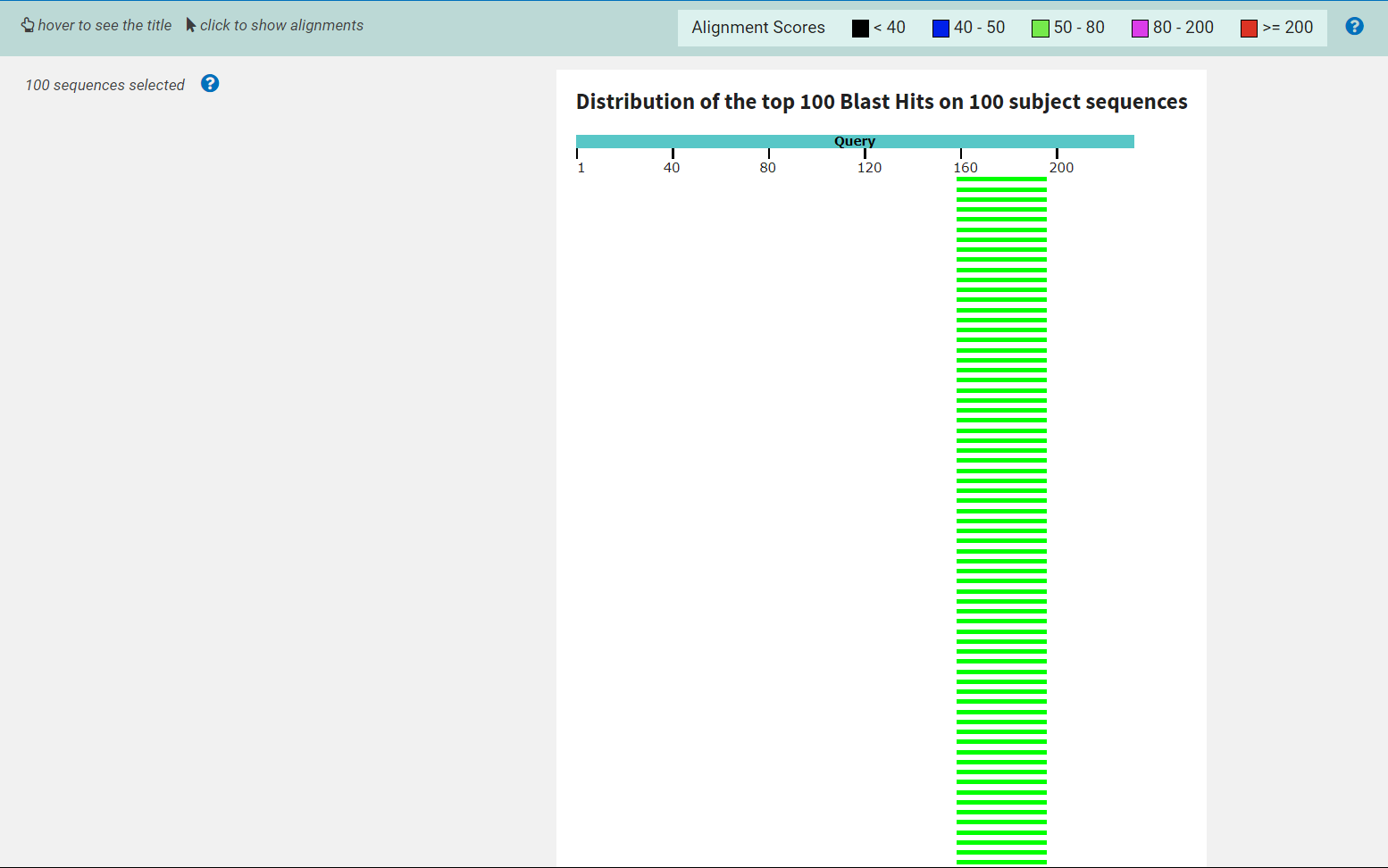
***Figure 1b:*** *Image of the results from blastn for Outbreak 14 in a graphical notation.*

An accompanying text file (./SEQ1/blastn/GM6THMYG016-Alignment.txt) with the whole data of the *blastn* results has been provided.

### ### *Megablast*



***Figure 2a:*** *Image of the results from megablast for Outbreak 14.*

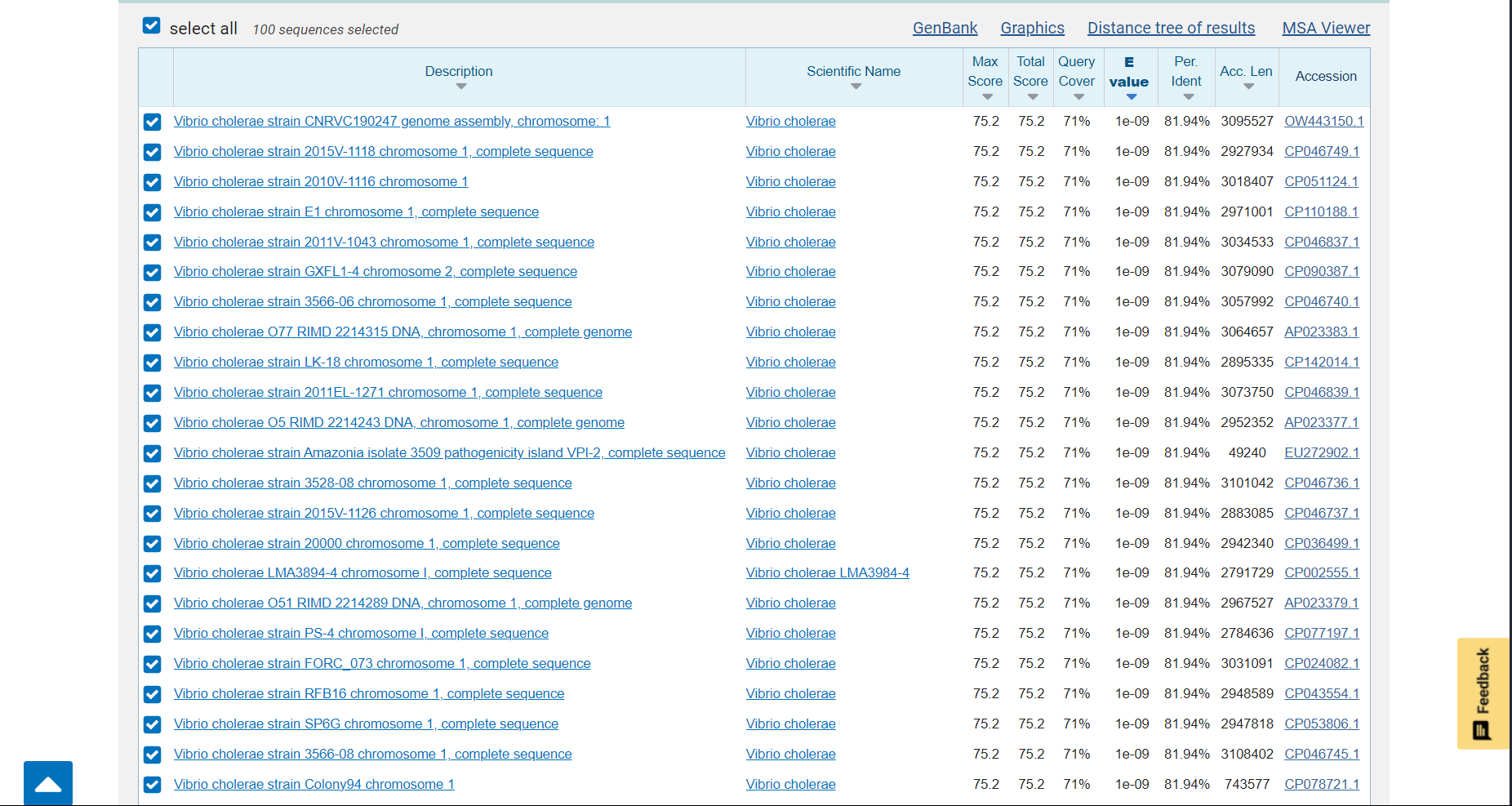


***Figure 2b:*** *Image of the results from megablast for Outbreak 14 in a graphical notation.*

An accompanying text file (./SEQ1/megablast/GM6TE49U016-Alignment.txt) with the whole data of the *megablast* results has been provided.

## ## Outbreak 15

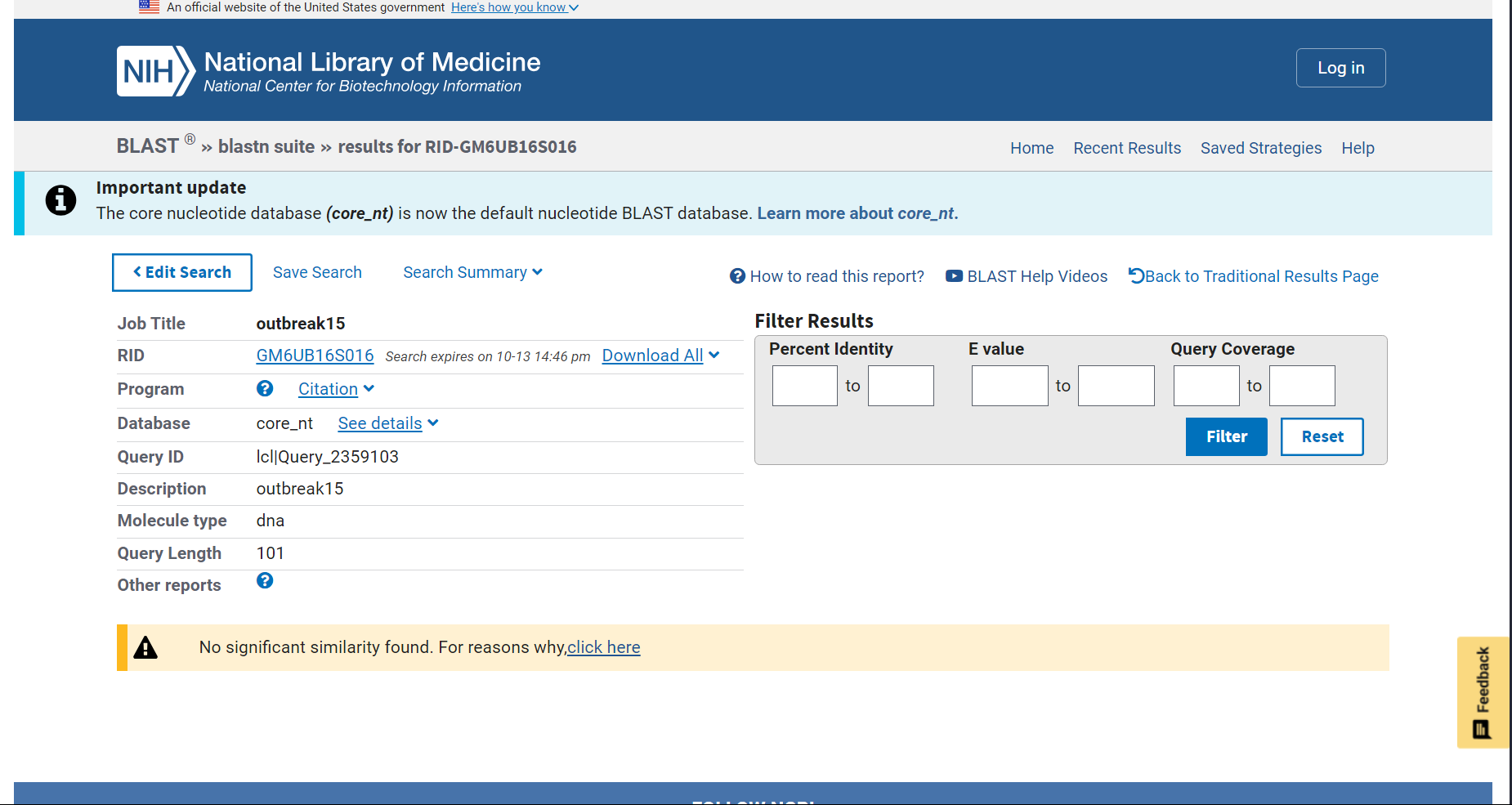
### ### *Blastn*



***Figure 3:*** *Image of the results from blastn for Outbreak 15.*

An accompanying text file (./SEQ2/blastn/GM6UGYYA013-Alignment.txt) with the whole data of the *blastn* results has been provided.

### ### *Megablast*



***Figure 4:*** *Image of the results from megablast for Outbreak 15.*

No accompanying text file can be provided for the *megablast* result for the results are empty.

# # Conclusions

Conclusions are founded primarily on the ***E-values*** of results (A lower ***E-value*** is indicative of a closer match).

## ## Differences in the algorithms

While conducting the analysis, it was obvious that *blastn* and *megablast* function differently in a way that leads to non-convergent results.

First, the *sensitivity* of each differs, in this case a little bit. This is evident by the E-***values*** of the *hits* provided, which differ between the *blast* methods. Secondly, the organisms found are not common and this difference is much more obvious and profound between the two *blast* methodologies. Moreso, the *megablast's* methodology in the case of Outbreak 15, is such that no *hits* are provided for the input sequence. However *megablast* is implemented, it would seem that there is a \**hard limit*\* on the similarity acceptable to constitute a result (or *hit*).

On these perceived differences, the listed descriptions of the two methods are:

*Megablast* is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast (Taken from the information provided directly on the https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\_TYPE=BlastSearch&LINK\_LOC=blasthome page).

*BlastN* is slow but allows a word-size down to seven bases (Taken from the information provided directly on the https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\_TYPE=BlastSearch&LINK\_LOC=blasthome page).

With these notes from the original authors (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\_TYPE=BlastSearch&LINK\_LOC=blasthome page) it would seem like the perceived differences are justified, for they interpret *megablast’s* results in a similar way. Put simply "A closely related sequence to the query", "the target percent identity is 95% or more" (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\_TYPE=BlastSearch&LINK\_LOC=blasthome) does not describe our data because our data is "DNA molecules are very short, and have been partially degraded.", "the sequence is sprinkled with Ns, “N” stands for “nucleotide” and means that the nucleotide at that position could not be determined." (Requirements). In a similar way the notes imply the coincidal results from *blastn* to ours.

## ## Outbreak 14

Based on the previous section (## Differences in the algorithms) we could disregard results from the *megablast* part of each sequence, we will not however. Reasons for this will be obvious further in the document. First, we should compare the organisms in the *hits* from both the *blast* and *megablast* to those listed as *bioterror* agents (<https://emergency.cdc.gov/agent/agentlist.asp>). This list is {*E. coli* O157:H7, *Rickettsia prowazekii*}. Then, one by one these sequences should be eliminated due to other variables that invalidate them in the *blast* results. It is at this point that we must notice the distribution of the *hits* in the results. The sequence seems tampered with, meaning that the parts of the sequence that map to different organisms in the *blast hits* are not from the same segment of the Outbreak 14 sequence. This is shown in Figure 1b, Figure 2b. This seems indicative that multiple pieces of genomes from organisms have been sewn together in the case of Outbreak 14. Having recognized this, the result is that we cannot focus on finding just one *hit* from the query data, it is now in the realm of possibility that the sequence of Outbreak 14 contains aspects of *bioterror agents* in small parts dispersed within its sequence.

We begin the search from the *megablast* results, searching for the case of {*E. coli* O157:H7}, the possible *bioterror agent* in the list of *Union hits* between the *blast* results and the listing at <https://emergency.cdc.gov/agent/agentlist.asp>. It is the case that we cannot say this is a sign of *bioterror* in the mall because other non-*bioterror* strains of *E. coli* all have the exact same sequence that maps onto Outbreak 14. This is exemplified if we take just one of the *E. coli* sequences in this region and do a *blast* (both *blastn, megablast)* on the region that Outbreak 14 maps to.

A screenshot of a computer code

Description automatically generated

***Figure 5a:*** *blastn result for the segment of E. coli that Outbreak 14 maps to.*

A screenshot of a computer

Description automatically generated

***Figure 5b:*** *megablast result for the segment of E. coli that Outbreak 14 maps to.*

From Figure 5a, Figure 5b, we see that the part of the *E.* coli genome that has a hit with Outbreak 14 is one that all *E. coli* share with a 100% coverage and 100% identity, meaning that strain {O157:H7} is no more likely to be the specific reason for the results in Figure 1a, 1b than any other *E.* coli strain. Thus, eliminating {*E.* coli O157:H7} from further consideration and thus concluding the results from the *megablast* of Outbreak 14 and figures 2a, 2b.

Onto *Rickettsia prowazekii*. Considering that the list of *bioterror* agents (<https://emergency.cdc.gov/agent/agentlist.asp>) does not delimitate between different strains of *Rickettsia prowazekii*, and the number of *hits* from *blast* for sequences belonging to *Rickettsia prowazekii* with good (relatively low) ***E-values*** in the *blastn* results, we can say that *Rickettsia prowazekii* is a sign of *bioterror* from the mall incident (Outbreak 14). All of this uses the relatively *bad* (high) ***E-value*** (statistical significance starts from ***E-values*** of 10-5 and below (González *et. al.,* 2019)) of [10-10, 10-5]. The ***E-value’s*** statistical significance is *subjective* in that you can interpret it however you like, however, there is an agreed upon standard for what constitutes statistical significance (González *et. al.,* 2019). Considering our sequences are “small” and “degraded” (Requirements), a worse (higher) ***E-value*** is more acceptable. Thus, these results will still be considered as evidence of bioterrorism.

Synthetic construct Note: In the *blastn* data the first, and best *hits* for Outbreak 14 are listed as “Synthetic construct”. These have the best ***E-values*** of all the results and map onto their own segment of the sequence of Outbreak 14 (Figures 1a, 1b). These segments are not listed as *bioterror* agents in the listing (<https://emergency.cdc.gov/agent/agentlist.asp>) but their peculiar origin means that they too should be evaluated for potential harmful effects that could constitute as *bioterrorism*. This is a sign of the sequences isolated being tampered with from human intervention and artificial construction. A further analysis should be performed on the segments that map onto the synthetic construct in the lab.

In total we will conclude that Outbreak 14 has signs of *bioterrorism* because of the circumstances regarding sequences from *Rickettsia prowazekii* in the *blast* results.

## ## Outbreak 15

Considering the results of the *megablast* only the results of the *blastn* search can be used for the identification of Outbreak 15 (Figure 3). From Figure 3, we compare the *hits* to the listings of *bioterror* agents from (https://emergency.cdc.gov/agent/agentlist.asp) and find that there is an overlap. Specifically, the organisms belonging to the *Vibrio cholerae* species, are listed as bioterror agents and are found in the *hits* for the sequence provided for Outbreak 15. Of these *hits* the most substantial ones are from the *Vibrio cholerae* species, which are also part of the listing at (https://emergency.cdc.gov/agent/agentlist.asp). The query coverage and distribution (Relative information in the ./SEQ2/blastn/GM6UGYYA013-Alignment.txt file from the alignments at the end of the file) are continuous and cover a large part of Outbreak 15’s sequence. Thus, signs of *bioterrorism* are found in Outbreak 15. All of this uses the relatively *bad* (high) ***E-value*** (statistical significance starts from ***E-values*** of 10-5 and below (González *et. al.,* 2019)) of [10-9, 10-8]. The ***E-value’s*** statistical significance is *subjective* in that you can interpret it however you like, however, there is an agreed upon standard for what constitutes statistical significance (González *et. al.,* 2019). Considering our sequences are “small” and “degraded” (Requirements), a worse (higher) ***E-value*** is more acceptable. Thus, these results will still be considered as evidence of bioterrorism.

## ## Final statement

Individuals caught in the incident should go to a diagnostic center and get checked for *Rickettsia prowazekii* and *Vibrio cholerae*. Individuals may experience:

* Watery diarrhea, vomiting, rapid heart rate, loss of skin elasticity, low blood pressure, thirst, and muscle cramps, kidney failure and possibly a coma.
* Fever, chills, headaches, rapid breathing, a rash, a cough, nausea.

# # Citations

<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome>

González-Pech, R. A., Stephens, T. G., & Chan, C. X. (2019). Commonly misunderstood parameters of NCBI BLAST and important considerations for users. *Bioinformatics (Oxford, England)*, *35*(15), 2697–2698. https://doi.org/10.1093/bioinformatics/bty1018

<https://emergency.cdc.gov/agent/agentlist.asp>

<https://www.ncbi.nlm.nih.gov/datasets/genome/>