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

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.

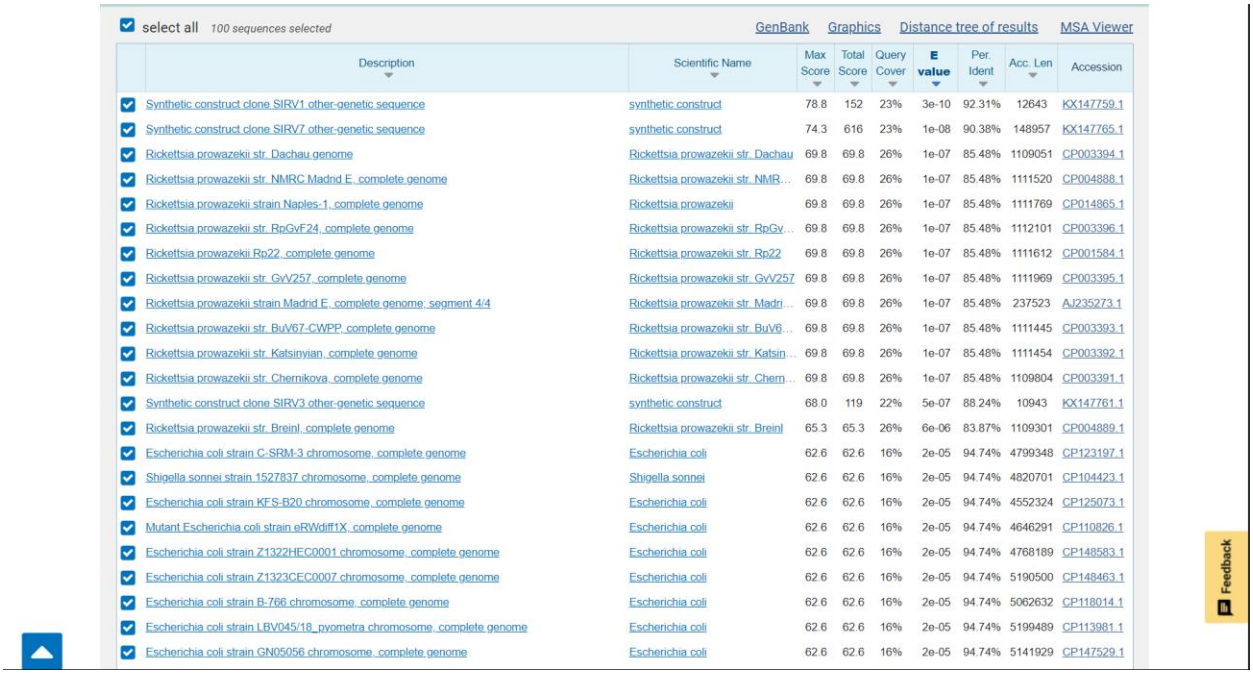
SEND SOLUTIONS (for M.Reczko exercises) ONLY TO:

mareczko@di.uoa.gr

Analysis

Outbreak 14

Blastn



select all 100 sequences selected — GenBank | Graphics | Distance tree of results | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Synthetic construct clone SIRV1 other-genetic sequence	synthetic construct	78.8	152	23%	3e-10	92.31%	12643	KX147759.1
Synthetic construct clone SIRV7 other-genetic sequence	synthetic construct	74.3	616	23%	1e-08	90.38%	148957	KX147765.1
Rickettsia prowazekii str. Dachau genome	Rickettsia prowazekii str. Dachau	69.8	69.8	26%	1e-07	85.48%	1109051	CP003394.1
Rickettsia prowazekii str. NMRC Madrid E, complete genome	Rickettsia prowazekii str. NMR…	69.8	69.8	26%	1e-07	85.48%	1111520	CP004888.1
Rickettsia prowazekii strain Naples-1, complete genome	Rickettsia prowazekii	69.8	69.8	26%	1e-07	85.48%	1111769	CP014865.1
Rickettsia prowazekii str. RpGvF24, complete genome	Rickettsia prowazekii str. RpGv…	69.8	69.8	26%	1e-07	85.48%	1112101	CP003396.1
Rickettsia prowazekii Rp22, complete genome	Rickettsia prowazekii str. Rp22	69.8	69.8	26%	1e-07	85.48%	1111612	CP001584.1
Rickettsia prowazekii str. GvV257, complete genome	Rickettsia prowazekii str. GvV257	69.8	69.8	26%	1e-07	85.48%	1111969	CP003395.1
Rickettsia prowazekii strain Madrid E, complete genome; segment 4/4	Rickettsia prowazekii str. Madri…	69.8	69.8	26%	1e-07	85.48%	237523	AJ235273.1
Rickettsia prowazekii str. BuV67-CWPP, complete genome	Rickettsia prowazekii str. BuV6…	69.8	69.8	26%	1e-07	85.48%	1111445	CP003393.1
Rickettsia prowazekii str. Katsinyian, complete genome	Rickettsia prowazekii str. Katsin…	69.8	69.8	26%	1e-07	85.48%	1111454	CP003392.1
Rickettsia prowazekii str. Chernikova, complete genome	Rickettsia prowazekii str. Chern…	69.8	69.8	26%	1e-07	85.48%	1109804	CP003391.1
Synthetic construct clone SIRV3 other-genetic sequence	synthetic construct	68.0	119	22%	5e-07	88.24%	10943	KX147761.1
Rickettsia prowazekii str. Breinl, complete genome	Rickettsia prowazekii str. Breinl	65.3	65.3	26%	6e-06	83.87%	1109301	CP004889.1
Escherichia coli strain C-SRM-3 chromosome, complete genome	Escherichia coli	62.6	62.6	16%	2e-05	94.74%	4799348	CP123197.1
Shigella sonnei strain 1527837 chromosome, complete genome	Shigella sonnei	62.6	62.6	16%	2e-05	94.74%	4820701	CP104423.1
Escherichia coli strain KFS-B20 chromosome, complete genome	Escherichia coli	62.6	62.6	16%	2e-05	94.74%	4552324	CP125073.1
Mutant Escherichia coli strain eRWdiff1X, complete genome	Escherichia coli	62.6	62.6	16%	2e-05	94.74%	4646291	CP110826.1
Escherichia coli strain Z1322HEC0001 chromosome, complete genome	Escherichia coli	62.6	62.6	16%	2e-05	94.74%	4768189	CP148583.1
Escherichia coli strain Z1323CEC0007 chromosome, complete genome	Escherichia coli	62.6	62.6	16%	2e-05	94.74%	5190500	CP148463.1
Escherichia coli strain B-766 chromosome, complete genome	Escherichia coli	62.6	62.6	16%	2e-05	94.74%	5062632	CP118014.1
Escherichia coli strain LBV045/18_pyometra chromosome, complete genome	Escherichia coli	62.6	62.6	16%	2e-05	94.74%	5199489	CP113981.1
Escherichia coli strain GN05056 chromosome, complete genome	Escherichia coli	62.6	62.6	16%	2e-05	94.74%	5141929	CP147529.1

Feedback

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

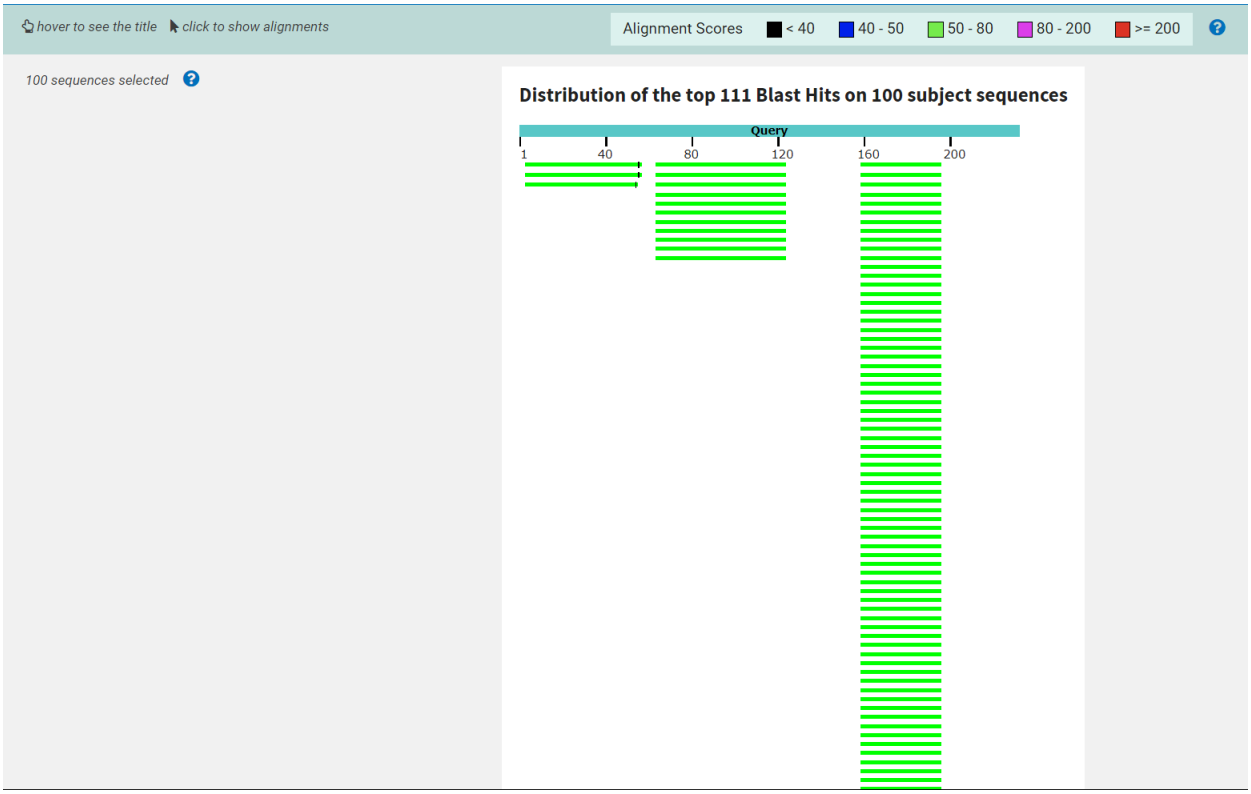
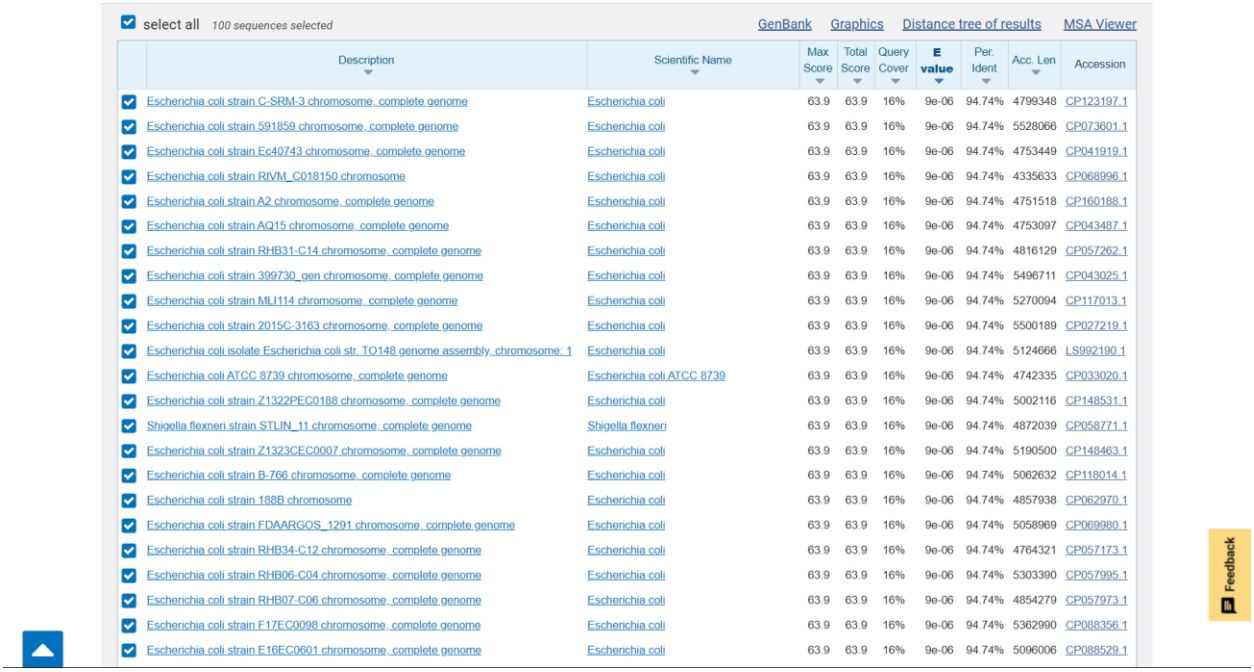


Figure 1b: *Image of the results from blastn for Outbreak 14 in a graphical notation.*

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

Megablast



select all 100 sequences selected — GenBank | Graphics | Distance tree of results | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Escherichia coli strain C-SRM-3 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4799348	CP123197.1
Escherichia coli strain 591859 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5528066	CP073601.1
Escherichia coli strain Ec40743 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4753449	CP041919.1
Escherichia coli strain RIVM_C018150 chromosome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4335633	CP068996.1
Escherichia coli strain A2 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4751518	CP160188.1
Escherichia coli strain AQ15 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4753097	CP043487.1
Escherichia coli strain RHB31-C14 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4816129	CP057262.1
Escherichia coli strain 399730_gen chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5496711	CP043025.1
Escherichia coli strain MLI114 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5270094	CP117013.1
Escherichia coli strain 2015C-3163 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5500189	CP027219.1
Escherichia coli isolate Escherichia coli str. TO148 genome assembly, chromosome: 1	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5124666	LS992190.1
Escherichia coli ATCC 8739 chromosome, complete genome	Escherichia coli ATCC 8739	63.9	63.9	16%	9e-06	94.74%	4742335	CP033020.1
Escherichia coli strain Z1322PEC0188 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5002116	CP148531.1
Shigella flexneri strain STLIN_11 chromosome, complete genome	Shigella flexneri	63.9	63.9	16%	9e-06	94.74%	4872039	CP058771.1
Escherichia coli strain Z1323CEC0007 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5190500	CP148463.1
Escherichia coli strain B-766 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5062632	CP118014.1
Escherichia coli strain 188B chromosome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4857938	CP062970.1
Escherichia coli strain FDAARGOS_1291 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5058969	CP069980.1
Escherichia coli strain RHB34-C12 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4764321	CP057173.1
Escherichia coli strain RHB06-C04 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5303390	CP057995.1
Escherichia coli strain RHB07-C06 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	4854279	CP057973.1
Escherichia coli strain F17EC0098 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5362990	CP088356.1
Escherichia coli strain E16EC0601 chromosome, complete genome	Escherichia coli	63.9	63.9	16%	9e-06	94.74%	5096006	CP088529.1

Feedback

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

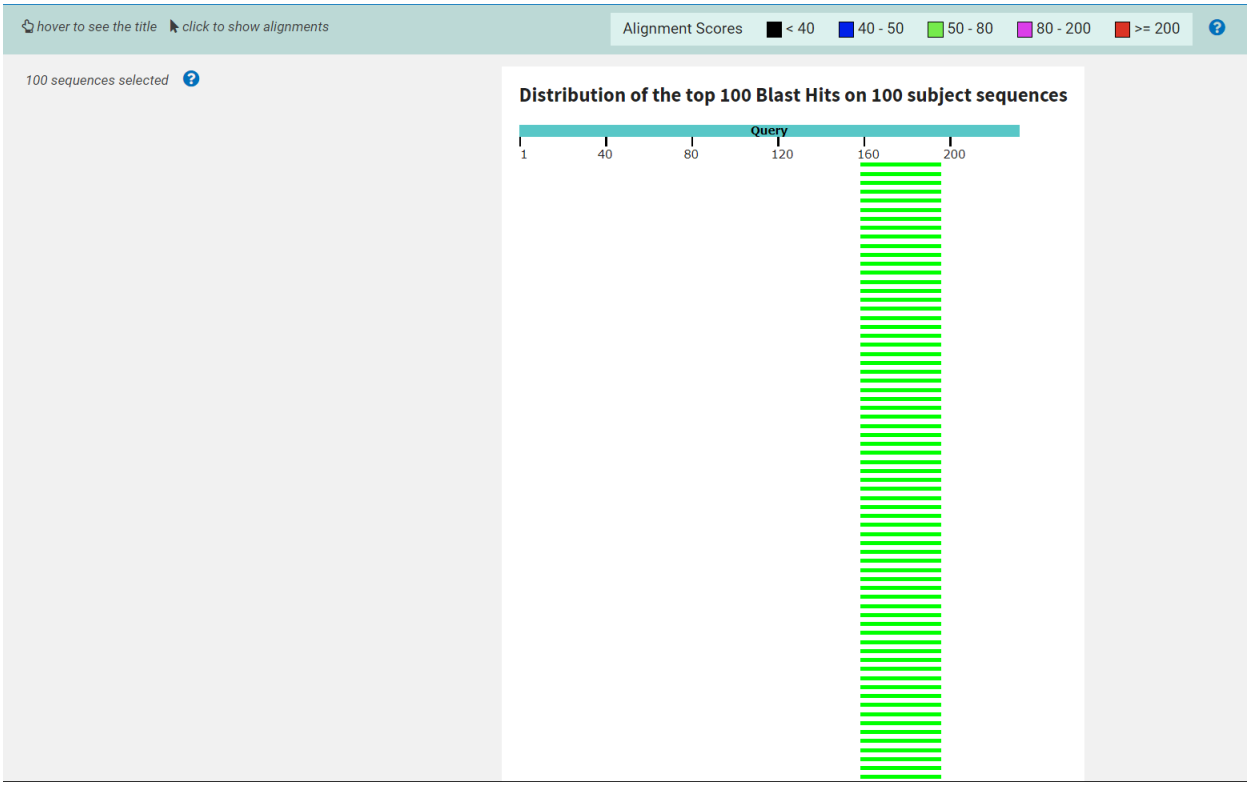
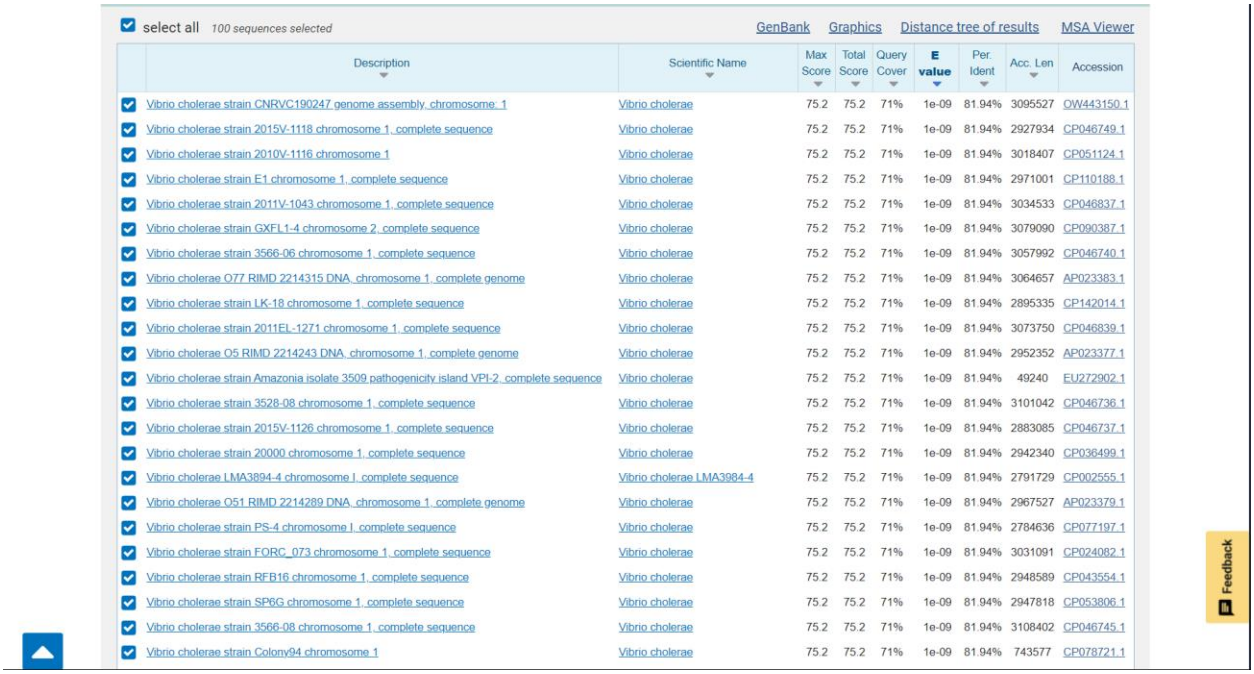


Figure 2b: *Image of the results from megablast for Outbreak 14 in a graphical notation.*

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

Outbreak 15

Blastn



select all 100 sequences selected — GenBank | Graphics | Distance tree of results | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Vibrio cholerae strain CNRVC190247 genome assembly, chromosome: 1	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3095527	OW443150.1
Vibrio cholerae strain 2015V-1118 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2927934	CP046749.1
Vibrio cholerae strain 2010V-1116 chromosome 1	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3018407	CP051124.1
Vibrio cholerae strain E1 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2971001	CP110188.1
Vibrio cholerae strain 2011V-1043 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3034533	CP046837.1
Vibrio cholerae strain GXFL1-4 chromosome 2, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3079090	CP090387.1
Vibrio cholerae strain 3566-06 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3057992	CP046740.1
Vibrio cholerae O77 RIMD 2214315 DNA, chromosome 1, complete genome	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3064657	AP023383.1
Vibrio cholerae strain LK-18 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2895335	CP142014.1
Vibrio cholerae strain 2011EL-1271 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3073750	CP046839.1
Vibrio cholerae O5 RIMD 2214243 DNA, chromosome 1, complete genome	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2952352	AP023377.1
Vibrio cholerae strain Amazonia isolate 3509 pathogenicity island VPI-2, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	49240	EU272902.1
Vibrio cholerae strain 3528-08 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3101042	CP046736.1
Vibrio cholerae strain 2015V-1126 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2883085	CP046737.1
Vibrio cholerae strain 20000 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2942340	CP036499.1
Vibrio cholerae LMA3894-4 chromosome I, complete sequence	Vibrio cholerae LMA3984-4	75.2	75.2	71%	1e-09	81.94%	2791729	CP002555.1
Vibrio cholerae O51 RIMD 2214289 DNA, chromosome 1, complete genome	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2967527	AP023379.1
Vibrio cholerae strain PS-4 chromosome I, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2784636	CP077197.1
Vibrio cholerae strain FORC_073 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3031091	CP024082.1
Vibrio cholerae strain RFB16 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2948589	CP043554.1
Vibrio cholerae strain SP6G chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	2947818	CP053806.1
Vibrio cholerae strain 3566-08 chromosome 1, complete sequence	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	3108402	CP046745.1
Vibrio cholerae strain Colony94 chromosome 1	Vibrio cholerae	75.2	75.2	71%	1e-09	81.94%	743577	CP078721.1

Feedback

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

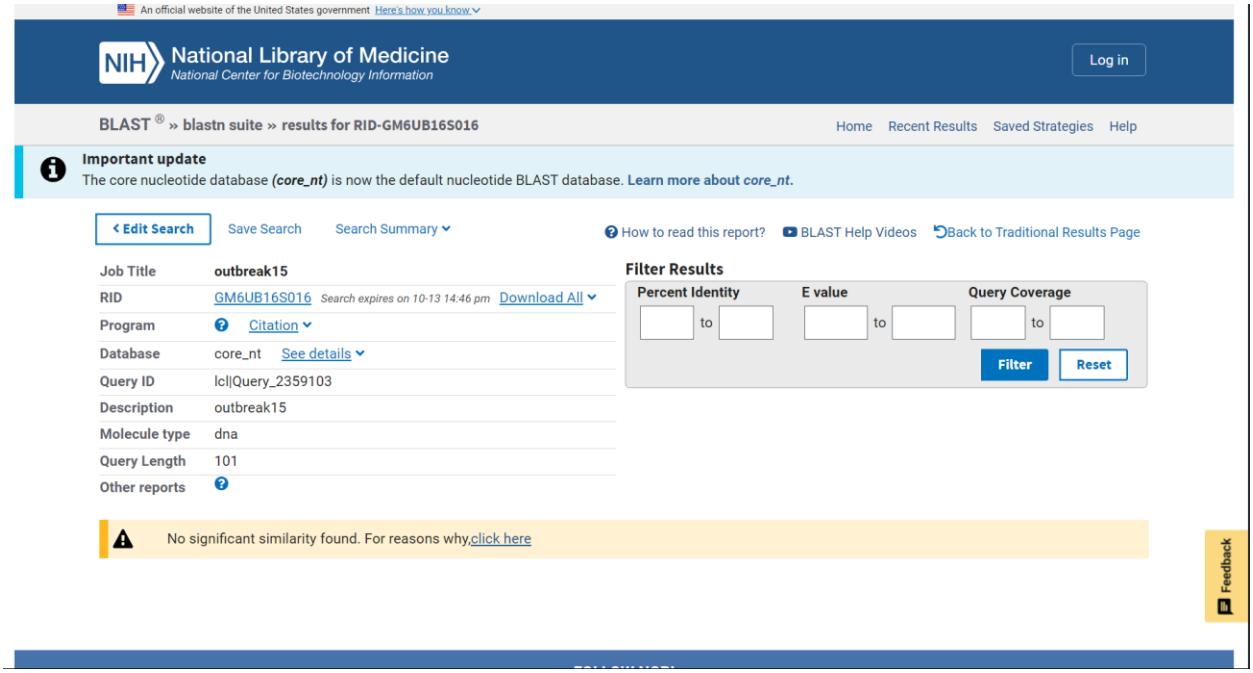


Figure 4: *Image of the results from megablast for Outbreak 15.*

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.

Escherichia coli strain C-SRM-3 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4799348	CP123197.1
Shigella sonnei strain 1527837 chromosome, complete genome	Shigella sonnei	69.8	69.8	100%	4e-09	100.00%	4820701	CP104423.1
Escherichia coli strain KFS-B20 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4552324	CP125073.1
Escherichia coli strain STEC506 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4966697	CP061239.1
Mutant Escherichia coli strain eRWdiff1X, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4646291	CP110826.1
Escherichia coli strain Z1322HEC0001 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4768189	CP148583.1
Escherichia coli strain Z1323CEC0007 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	5190500	CP148463.1
Escherichia coli strain B-766 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	5062632	CP118014.1
Escherichia coli strain C16EC0166 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4999191	CP088681.1
Escherichia coli strain LBV045/18_pyometra chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	5199489	CP113981.1
Escherichia coli strain GN05056 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	5141929	CP147529.1
Escherichia coli strain 2A chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4833996	CP146939.1
Escherichia coli strain TUM2367 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	5023218	CP135697.1
Escherichia coli O26:H11 EH031 DNA, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	5730399	AP027162.1
Escherichia coli strain B-DidR chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4611875	CP115968.1
Escherichia coli strain RHB13-SO-C03 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4859593	CP099225.1
Escherichia coli isolate CNR65D6 genome assembly, chromosome: 1	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	5175614	OU701452.1
Escherichia coli strain EC812A1 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4665801	CP116046.1
Escherichia coli strain S-P-C-029.01 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4641599	CP092702.1
Escherichia coli strain RHB21-E3-C02 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	5274828	CP099181.1
Escherichia coli strain Z0117EC0055 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4954543	CP098203.1
Escherichia coli strain ETEC4088 chromosome, complete genome	Escherichia coli	69.8	69.8	100%	4e-09	100.00%	4978151	CP122629.1

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

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Escherichia coli strain C-SRM-3 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4799348	CP123197.1
Escherichia coli strain 591859 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5528066	CP073601.1
Escherichia coli strain Ec40743 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4753449	CP041919.1
Escherichia coli strain RIVM_C018150 chromosome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4335633	CP068996.1
Escherichia coli strain A2 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4751518	CP160188.1
Escherichia coli strain AQ15 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4753097	CP043487.1
Escherichia coli strain RHB31-C14 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4816129	CP057262.1
Escherichia coli strain 399730_gen chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5496711	CP043025.1
Escherichia coli strain MLI114 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5270094	CP117013.1
Escherichia coli strain 2015C-3163 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5500189	CP027219.1
Escherichia coli isolate Escherichia coli str. TQ148 genome assembly, chromosome: 1	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5124666	LS992190.1
Escherichia coli ATCC 8739 chromosome, complete genome	Escherichia coli ATCC 8739	71.3	71.3	100%	2e-09	100.00%	4742335	CP033020.1
Escherichia coli strain Z1322PEC0188 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5002116	CP148531.1
Shigella flexneri strain STLIN_11 chromosome, complete genome	Shigella flexneri	71.3	71.3	100%	2e-09	100.00%	4872039	CP058771.1
Escherichia coli strain Z1323CEC0007 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5190500	CP148463.1
Escherichia coli strain B-766 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5062632	CP118014.1
Escherichia coli strain 188B chromosome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4857938	CP062970.1
Escherichia coli strain FDAARGOS_1291 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5058969	CP069980.1
Escherichia coli strain RHB34-C12 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4764321	CP057173.1
Escherichia coli strain RHB06-C04 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5303390	CP057995.1
Escherichia coli strain RHB07-C06 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	4854279	CP057973.1
Escherichia coli strain F17EC0098 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5362990	CP088356.1
Escherichia coli strain E16EC0601 chromosome, complete genome	Escherichia coli	71.3	71.3	100%	2e-09	100.00%	5096006	CP088529.1

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/