


[◀ Back to Week 1](#)[✕ Lessons](#)[Prev](#)[Next](#)

Review Your Peers: Application Challenge 1

Was due July 26, 11:59 PM PDT

Reviews 3 left to complete

 It looks like this is your first peer-graded assignment. [Learn more](#)





e.coli x



by Michael Gotama

July 17, 2017

 like  Flag this submission

PROMPT

But before you embark on analyzing the *E. coli* X strain, we will take the time to explain how a harmless bacterium can become pathogenic in the first place.

How can a bacterium become pathogenic?

There are more than 700 known infectious subspecies, or strains, of *E. coli* (see FAQ: Classification of pathogenic *E. coli* strains). *E. coli* pathogenicity is determined by **virulence factors**, various compounds that the bacterium produces to colonize the host and evade or inhibit the host's immune system. Bacteria possess a wide array of virulence factors that may be encoded either in the bacterial chromosome or in extra-chromosomal genetic elements called **plasmids**, circular self-replicating mini-chromosomes that co-exist with the bacterial genome (Figure below). Plasmids, which are passed on to daughter cells

Help Center

during replication, often provide selective advantages to an organism. For example, they may carry genes for antibiotic resistance, or pathways allowing bacteria to use additional energy sources or to survive in harsh environments.

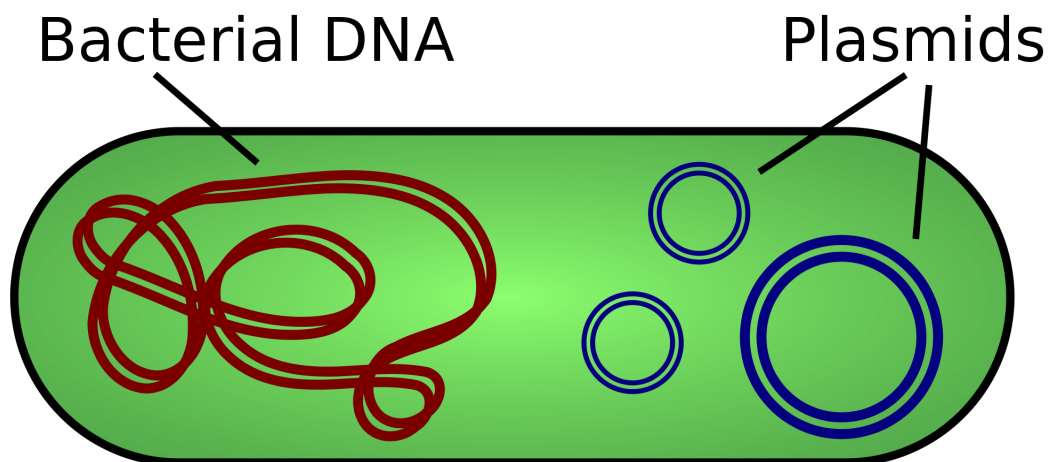


Figure: Plasmids in a bacterial cell. Courtesy: Spaully, Wikimedia Commons user.

Many pathogenic bacteria secrete **toxins**, substances (small molecules, peptides, or proteins) that may inhibit cellular functions in the host. Examples include the **tetanus toxin** secreted by *Clostridium tetani*, the **botulinum toxin** secreted by *Clostridium botulinum*, and the **anthrax toxin** produced by *Bacillus anthracis*. Several strains of *E. coli* and *Shigella* secrete **Shiga toxins**, which are encoded by **Stx** genes. Shiga toxins are named for Kiyoshi Shiga, who described the bacterial origin of a dysentery outbreak in Japan in 1897 during which nearly 30,000 people died. These toxins cause bleeding by breaking down the lining of the colon, and they can lead to HUS if they reach the kidneys. Shiga toxins attack highly specific receptors on the surface of human cells, and so species that do not have this receptor, such as cows, may harbor toxigenic bacteria without any ill effects.

Another type of pathogenic agents are **phages**, viruses that cannot replicate on their own and must infect bacteria to do so. Many phages are shaped like lunar landers (Figure below), a design that helps them land on the cell wall of a bacterium and transmit their own genome (called a **prophage**) into the bacterial genome, so that when the bacterial DNA replicates, it creates new copies of the phage as well.

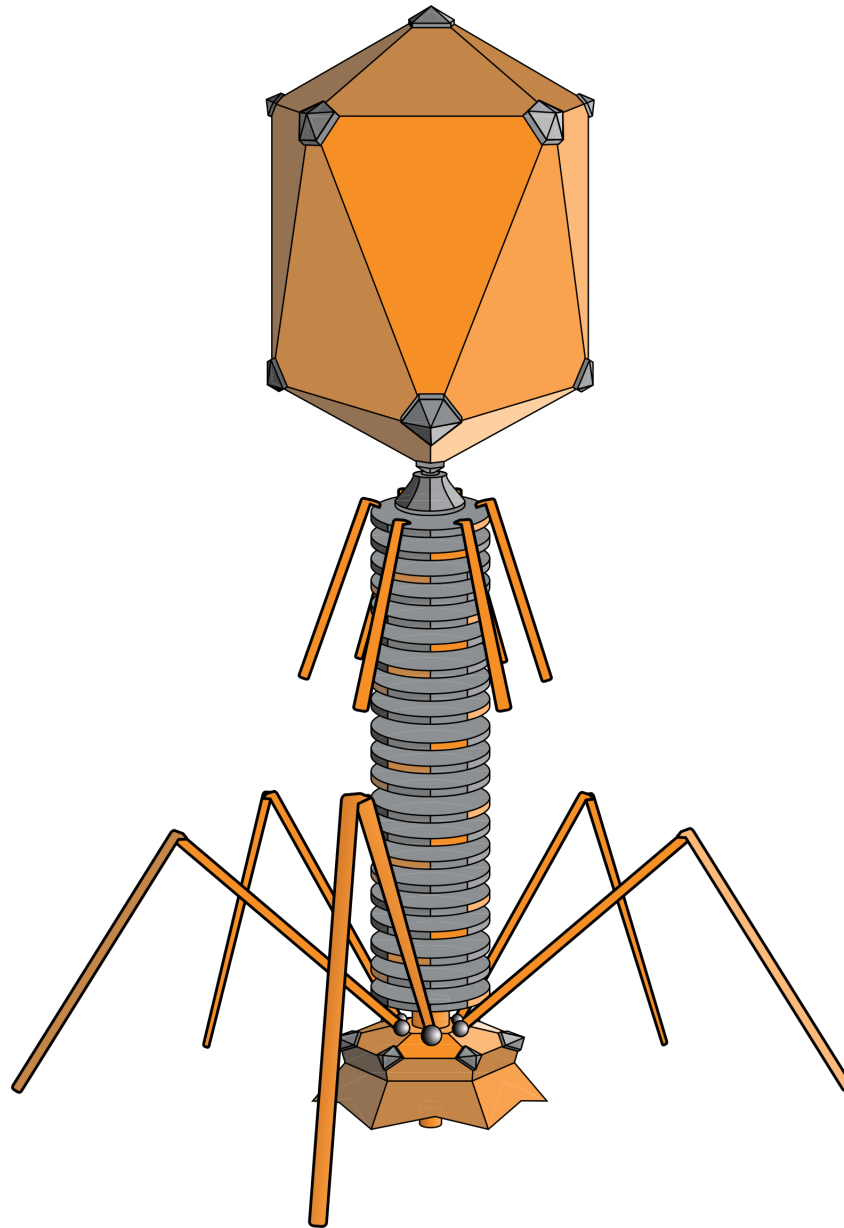


Figure: The structure of a T4 bacteriophage. Courtesy: Adenosine, Wikimedia Commons user.

Sometimes, the prophage is replicated as a plasmid; in other cases, prophage genes encode recombinases, enzymes that can catalyze DNA exchange reactions between short (30–200 nucleotides) similar sequences. During this process, called **site-specific recombination**, prophage literally "glued" into host DNA.

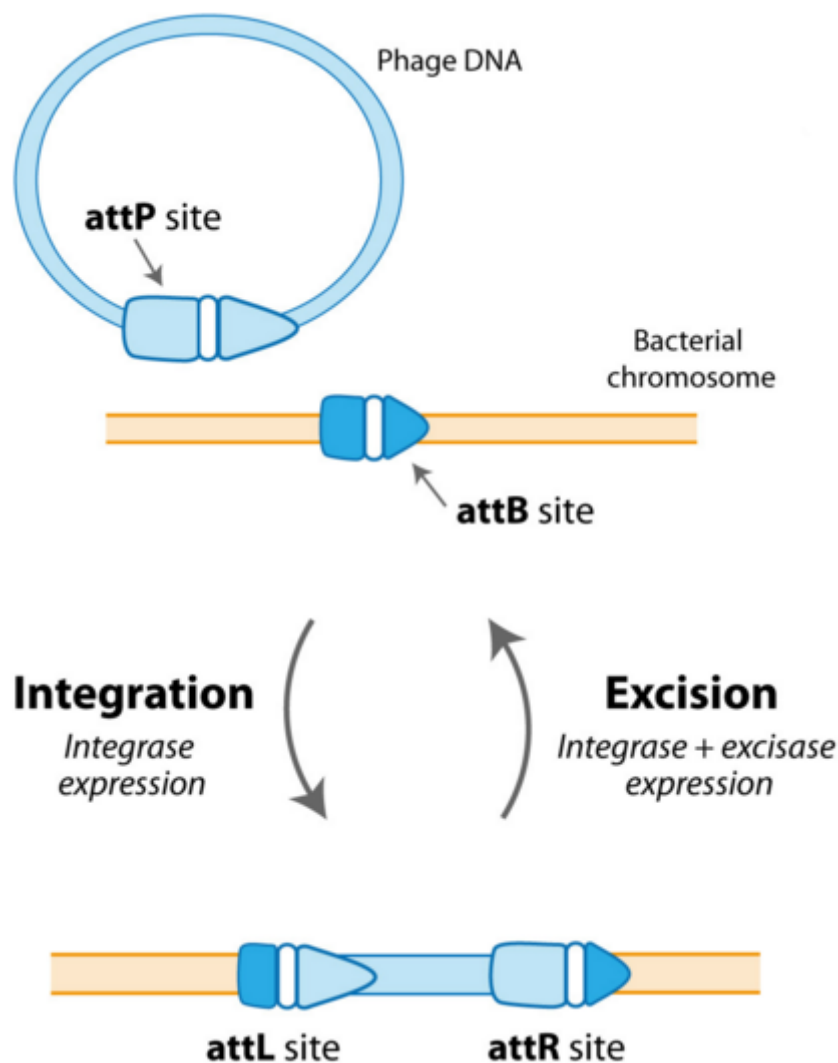


Figure: Phage DNA integration in a bacterial chromosome via recombination between the phage attP and the chromosomal attB recombination sites. Source: Memo-Cell project.

Phages offer an example of a phenomenon that was first witnessed in 1951 when Victor Freeman transferred a viral gene into the bacterium *C. diphtheriae* and transformed an otherwise non-virulent strain into a virulent one. Such **horizontal genetic transfer (HGT)**, or the transfer of genes between two organisms, is in contrast to the usual **vertical gene transfer** from a cell to its daughter cells during division. For that matter, phages do not always harm the host; by inserting their DNA into the host bacterium, they may actually provide the host with benefits by equipping its genome with new functions. For example, phages can transform harmless strains of *Vibrio cholerae* into the virulent ones that cause cholera. Furthermore, HGT is not limited to transfer between organisms of different types, as bacteria exchange mobile elements such as phages and plasmids to neighboring cells.

Sometimes, genes have been transferred horizontally and become a part of the host's DNA, being passed down to offspring. For example, the cellular organelle called the mitochondrion has its own DNA, and biologists have proposed that the

mitochondrion is the descendant of a group of bacteria called **alphaproteobacteria**, which were engulfed long ago by eukaryotic cell but not digested. Because egg cells have mitochondria but sperm cells do not, you inherited your “mitochondrial genome” from your mother, who inherited it from her mother, and so on back to the dawn of civilization. A similar example of horizontal gene transfer in plants is the chloroplast, which is a descendant of a cyanobacterium, acquired long ago by a plant ancestor to facilitate photosynthesis.

Assembling and Annotating the *E. coli* X genome

We will use the SPAdes assembler (Bankevich et al., 2012) to assemble the *E. coli* X genome from reads. A **contig** is a “contiguous” segment of the genome that has been reconstructed by an assembly algorithm. In addition to contigs, SPAdes uses information about the distances between reads within read-pairs (called **insert size**) to combine contigs into ordered collections of adjacent contigs called **scaffolds**. For example, as illustrated in the figure below, if one read in a read-pair appears in Contig 1, and the other appears in Contig 2, then we may infer that Contig 1 and Contig 2 are neighbors in the same scaffold. Genome assembly programs often fill a gap in a scaffold of length *m* by a sequence of *m* occurrences of “N” (a placeholder for unknown nucleotides).

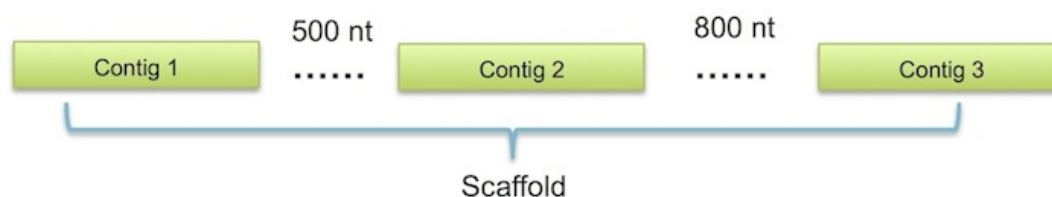


Figure: Three contigs combined into a single scaffold separated by two gaps of length 500 nt and 800 nt.

In modern DNA sequencing projects, DNA fragments are sequenced from both the 5' and 3' ends, giving rise to **paired reads** separated by some insert size. The **forward** and **reverse read** in a paired read are generated from the forward and reverse strand, respectively. A collection of paired reads is called a **sequencing library**; whereas most paired libraries generate reads with insert sizes below 1 kb, libraries with longer (2 kb-10 kb) insert sizes are called **mate-pairs**. (See FAQ: Paired-end and mate-pair reads.) Many sequencing projects, generate several libraries of paired reads with different insert sizes; for example, the sequencing project at the center of this Application Challenge employs three libraries of insert lengths 270, 2000, and 6000 nucleotides. We will not delve into the algorithmic details of how to produce a single assembly from multiple libraries with different insert sizes.

The advantage of using multiple libraries is that libraries with small insert sizes are better suited for resolving short repeats, whereas libraries with larger insert sizes are better suited for resolving long repeats. For example, all bacterial genomes contain **ribosomal operons**, which are often repeated more than six times and are approximately 5000 nucleotides long. Thus, to resolve ribosomal operons, we need mate-pair libraries with insert sizes exceeding 5000 nt.

Once we have sequenced a genome, there are many metrics for assessing the resulting assembly. (see FAQ: What are quality scores?). We will use QUAST (Gurevich et al., 2013), which takes many metrics into account to evaluate the overall quality of an assembly.

After assembling a genome, biologists **annotate** the genome, searching for genes and other important regions. After identifying putative genes, biologists perform **functional gene annotation** to determine their functions. The **comparative genomics** approach to gene annotation is based on the (not always ideal) assumption that similar genes in different organisms perform similar functions. Finding similar genes can be accomplished by running BLAST or by matching putative proteins against protein domains in the Pfam database.

Part 1: Assembling the *E. coli* X Genome

We will first use SPAdes to assemble a single library of sequencing (paired end) reads from *E. coli* X.

Important note: In the Application Challenges accompanying this capstone, we have already run the software that you will need to complete our analysis, and so we will often provide you with the results of running this software. We did this to save you time and allow you to complete each Application Challenge in one sitting; many of the programs we will encounter take a long time to run on real datasets -- as much as ten hours for some programs. (It also prevents a small error in running a tool from costing you points on multiple questions.) If you are interested in seeing the steps that are taken to run software on BaseSpace, we will provide a sequence of "walkthroughs" illustrating these steps.

So, after creating an account and logging into BaseSpace (<https://basespace.illumina.com>), please visit the following link, where we have shared all of the necessary data with you:
<https://basespace.illumina.com/s/6ZXyPRU9NUOp>

Under "Analyses", click on "SPAdes Genome Assembler (SRR292678)". SRR292678 is the ID of the sequencing library.

Question 1 (warm-up): Select "Analysis info" in the left menu and attach a screenshot of the corresponding page (no need to include the "logs" in the image).

(Note: To see how to run SPAdes yourself, see Walkthrough: How to run SPAdes.

The screenshot shows the BaseSpace Sequence Hub interface. The top navigation bar includes 'DASHBOARD', 'PREP', 'RUNS', 'PROJECTS', 'APPS', and 'PUBLIC DATA'. The user is logged in as 'Michael Gotam'. The main content area displays 'Analysis Info' for the 'SPAdes Genome Assembler (SRR292678)' project. The status is 'Complete' with the message 'Application completed successfully'.

Analysis Info	
Name	SPAdes Genome Assembler (SRR292678)
Application	SPAdes Genome Assembler Version: 3.6.0
Date Started	Sunday, July 10 2016 7:06:13 AM
Date Completed	Sunday, July 10 2016 7:52:11 AM
Duration	45 minutes 58 seconds
Compute Charge	0.00 iCredits
Session Type	Single Node
Size	10.46 MB
Status	Complete Application completed successfully

RUBRIC

Learner submission should resemble the following screenshot:

The screenshot shows the BaseSpace Sequence Hub interface. The top navigation bar includes 'DASHBOARD', 'PREP', 'RUNS', 'PROJECTS', 'APPS', and 'PUBLIC DATA'. The user is logged in as 'Mikhail Rayko'. The main content area displays 'Analysis Info' for the 'Capstone_assembly_test: SPAdes Genome Assembler 12/18/2015' project. The status is 'Complete' with the message 'Application completed successfully'.

Analysis Info	
NAME	SPAdes Genome Assembler 12/18/2015 12:47:56
APPLICATION	SPAdes Genome Assembler Version: 3.6.0
DATE STARTED	Thursday, December 17 2015 9:48:38 PM
DATE COMPLETED	Thursday, December 17 2015 10:11:31 PM
DURATION	22 minutes 53 seconds
SESSION TYPE	Single Node
SIZE	10.37 MB
STATUS	Complete Application completed successfully

Logs (last checked 10:15:59pm UTC)

Log Files

```

2015-12-17 21:51:34.987 [INFO] - Finished downloading Sample 'SRR292862' (Id: 31641610)
2015-12-17 21:51:34.990 [INFO] - Validating 'SRR292862_S1_L001_R2_001.fastq.g
  
```

- ☐ 0 pts
Answer is incorrect
- ☐ 1 pt
Answer is correct


Help Center

PROMPT


Evaluating assembly quality with QUASt


Now that SPAdes has finished running, we will assess the quality of the resulting assembly. There are many files in the SPAdes output, but for the subsequent analysis we will be interested only in the "contigs.fasta" and "scaffolds.fasta" files, which define contigs and scaffolds for the *E. coli* assembly, respectively.

We will use QUAST to evaluate the quality of our assembly. This program is incorporated into SPAdes on BaseSpace, so when SPAdes has completed, you can select this assembly in the project folder, and see the QUAST results by visiting "summary" under "Analysis reports". Alternatively, you can download the files contigs.fasta and scaffolds.fasta containing contig/scaffold information (under "Output Files") and use the online version of QUAST (quast.bioinf.spbau.ru).

 Analysis Info

 Inputs

 Output Files

 Analysis Reports

Summary

Consider the "contigs" column in the QUAST report. Note that all statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (the statistics "# contigs (≥ 0 bp)" and "Total length (≥ 0 bp)" refer to all contigs). Biologists are mainly interested in long contigs because short contigs often contain only gene fragments.

Question 2: With respect to contigs, what are N50 and the number of contigs of size ≥ 500 bp for the assembly?

111860, 210

RUBRIC

N50: **111,860**

Number of contigs of size ≥ 500 bp: **210**

☐ 0 pts

Both answers are incorrect

- ☐ 1 pt
One of the answers is correct
- ☐ 2 pts
Both answers are correct

PROMPT

Question 3: With respect to scaffolds, what are N50, the number of scaffolds of size ≥ 500 bp, and the average number of occurrences of "N" per 100 kb (rounded to the nearest integer)?

111860, 221, 34

RUBRIC

N50: **111,860**

Number of scaffolds of size ≥ 500 bp: **221**

"N" per 100 kb: **34**

- ☐ 0 pts
Both answers are incorrect
- ☐ 1 pt
One or two of the answers are correct
- ☐ 2 pts
All answers are correct

PROMPT

Impact of reads with large insert size

As mentioned in the introduction, there is an advantage in using multiple libraries with different insert sizes, since the library with a small insert size can resolve short repeats, whereas the library with a larger insert size can resolve longer

repeats. In particular, we are interested in how the quality of our assembly changes when we add mate-pair libraries.

This time, we will run SPAdes again by consolidating three libraries. Return to the project homepage by clicking the project name "E coli" in the top left (or by visiting the "Projects" page), and now select "SPAdes Genome Assembler (SRR292678, SRR292862, SRR292770)". The assembly report should be shown, but if not you can find it under "Analysis Reports" --> "Summary".

Question 4: With respect to scaffolds, what is N50, the number of scaffolds of size ≥ 500 bp, and the average number of occurrences of "N" per 100 kb (rounded to nearest integer) for this assembly? How did the quality of the assembly improve compared to the previous run of SPAdes?

Note: To see how to run SPAdes yourself, see Walkthrough: Running SPAdes with multiple libraries.

2815616, 90, 628. The N50 is increasing, hence the average length of scaffold is increase. This is better as the fragmentation is lesser (contig is longer), however the number of unknown base is also increase (need careful analysis).

RUBRIC

N50:

2,815,616 (previous answer was **159,018**, also acceptable)

Number of scaffolds of size ≥ 500 bp: **90** (previous answer was **118**, also acceptable)

"N" per 100 kb: **627.52** (previous answer was **33, 628** is also acceptable)

The statistics of scaffolds illustrates that using multiple libraries dramatically improve the contiguity of assembly as compared to using a single paired end libraries with a small insert size.

- ☐ 0 pts
Learner did not describe assembly comparison, and all values are incorrect
- ☐ 1 pt
Learner did not describe assembly comparison, or more than one value is incorrect

☐ 2 pts

Learner described assembly comparison, and all values are correct

PROMPT

Part 2: Annotation and analysis of the *E. coli* X genome

Now that we have sequenced the *E. coli* X genome, we will use Prokka (<http://www.vicbioinformatics.com/software/prokka.shtml>) for gene prediction and annotation. This tool identifies the coordinates of putative genes within contigs, and then uses BLAST for similarity-based annotation using all proteins from sequenced bacterial genomes in the [RefSeq](#) database.

We have run Prokka on the scaffolds.fasta file provided as part of the output of the most recent (three-library) assembly. You can find it by visiting the main "E coli" project page. Click on "Prokka Genome Annotation" to see the results.

Then, visit the "Output Files" page, select "scaffolds.gbk" from the output folder and store it on your computer, as we will use it later to compare *E. coli* X to a similar bacterium.

Question 5: Visit "Analysis Reports" --> "Summary". How many coding DNA sequences (CDS) did Prokka find?

Note: To see how to run Prokka yourself, check out Walkthrough: How to run Prokka.

5,064

RUBRIC

5,064 (previous answer was 5,228, also acceptable)

☐ 0 pts

Answer is incorrect

☐ 2 pts

Answer is correct

PROMPT

Finding the closest relative of *E. coli* X

Our goal is to find the known genome that is the most similar to the pathogenic strain (and infer properties of *E. coli* X from it). We could compare each contig in our assembly against the entire RefSeq database using BLAST, but this could take several hours depending on server workload. A more efficient approach is to select one important and evolutionarily conserved gene (**16S ribosomal RNA**) for comparison with all other sequenced genomes. The gene that we will use is **16S ribosomal RNA** (see [FAQ: "What is 16S ribosomal RNA?"](#)).

First, we need to locate 16S rRNA in the assembled *E. coli* X genome, for which we will use [RNAmmer](#). First, download the "scaffolds.fasta" file from the list of output files in the most recent run of SPAdes. Then, visit the RNAmmer website, upload "scaffolds.fasta", and press "Submit". The job will only take a few minutes to complete, so we will leave this task to you. When the job does finish, select "Download prediction result" → "Fasta" and copy the identified sequence of the 16S rRNA in the assembled genome.

As we mentioned before, rRNA genes in bacteria are typically organized in **ribosomal operons** – set of closely located genes that are activated together. Ribosomal RNA plays a crucial role in protein synthesis, and in order to achieve high growth rate bacteria often possess several copies of this operon. That is why you will probably get several matches here.

Question 6: What is the length of 16S rRNA in *E. coli* X? (You can use a text editor or online program to find the length of a string.)

1530

RUBRIC

1530

- ☐ 0 pts
Answer is incorrect
- ☐ 2 pts
Answer is correct

Help Center

PROMPT

We will now use BLAST to search for the genome in the **RefSeq** database with 16S rRNA that is most similar to the 16S rRNA that we just found. Open the NCBI BLAST homepage (<http://blast.ncbi.nlm.nih.gov>) and select "Nucleotide blast". To perform the search against complete genomes in the RefSeq database, select the "**Reference genomic sequences**" in the "**Database**" field. To restrict our search to only those genomes that were present in the GenBank database at the beginning of 2011, set the time range using parameter PDAT in the "Entrez Query" field:

1900/01/01:2011/01/01[PDAT]

Other parameters should be specified as default.

NCBI/ BLAST/ blastn suite **Standard Nucleotide BLAST**

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCTAACACATGCAAGTCGAACGGTAACA
GGAAACAGCTTGCTGTTTCCTGACGAGTGGCGGACGGTGAGTAATGCTGGGAACTGCCTGATG
GAGGGGATAACTACTGGAACGGTAGCTAATACCGCATAACGTCGAAGACCAAGAGGGGGACCT
TCGGGCTCTTGCCATCGGATGCCCAGATGGGATTAGCTTGTGGTGGGGTAACGGCTCACCAG
GCGACGATCCCTAGCTGGTCTGAGAGGATGACCCAGCCACACTGGAAGTGAAGACACGGTCCAGACTCC

Clear Query subrange [?](#)

From

To

Or, upload file No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database ☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):

Reference genomic sequences (refseq_genomic) [?](#)

Organism ☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to ☐ Sequences from type material

Entrez Query [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

When the search finishes after a few minutes, explore its results – we will use the closest relative to *E. coli* X as a reference genome. We will call this reference genome ***E. coli* A** for brevity.

Question 7: How many matches did you find for the rRNA of *E. coli* X in this closest genome (i.e. how many ribosomal operons there)?

7 ribosomal operons

RUBRIC

7 matches

- ☐ 0 pts
Value is incorrect
- ☐ 2 pts
Value is correct

PROMPT

Click on the "Sequence ID" link under the name of the identified reference in order to open its corresponding GenBank page. Download the genome sequence in FASTA format (in the right upper corner select "Send" → "Complete Record" → "File" → "Fasta", and save as "EcoliA.fasta").

Question 8: What is the accession number for the reference *E. coli* A strain identified in the GenBank database?

NC_011748.1

RUBRIC

NC_011748.1

(NC_011748 is also acceptable - .1 here is just a version number of the given nucleotide sequence.)

- ☐ 0 pts
Answer is incorrect
- ☐ 1 pt
Answer is correct

PROMPT

Comparing other regions of the *E. coli* X genome against the entire RefSeq database would tell us that *E. coli* A is its closest relative, but that *E. coli* X is nevertheless a distinct strain, one whose genome in early 2011 was unknown. The in-depth study of *E. coli* X began after the start of the outbreak; after you

complete this challenge, we encourage you can perform a search with assembled contigs, removing the restriction in the "Entrez Query" field, and explore the results on your own.

Identifying the genetic cause of HUS

Now that we have identified the closest relative of the strain that caused the outbreak, it makes sense to examine the original paper about this reference strain *E. coli* A (Mossoro et al., 2002). This paper states that *E. coli* A belongs to an enteroaggregative *E. coli* (EAEC) strain, harboring the pAA plasmid, which contains aggregative adherence fimbria (AAF) genes allowing bacteria to stick to cells in the intestine (see FAQ: Classification of pathogenic *E. coli* strains).

E. coli A was isolated in the Central African Republic from a stool sample obtained from a man with persistent diarrhea. However, in contrast to patients infected in the 2011 outbreak, this patient had no signs of blood in his stool! This fact suggests that *E. coli* X somehow gained an additional virulence factor over the previously identified *E. coli* A. But what is the virulence factor, and how did *E. coli* X obtain it?

To understand the genetic cause of HUS, we will perform a genome-wide comparison with *E. coli* A and will analyze the regions where these strains differ from each other. If we find a region where *E. coli* X encodes a new virulence factor or a new gene responsible for antibiotic resistance, it may shed light on the genetic cause of HUS.

We will use a program called Mauve, which visualizes an alignment as a series of conserved segments called Locally Collinear Blocks (LCBs), which are similar to syntenic blocks. Insertions and deletions in LCBs correspond to insertions and deletions in a bacterial chromosome. Separate unaligned regions that have no flanking regions from chromosomal DNA, on the other hand, may correspond to extrachromosomal elements such as plasmids. LCBs are indicated using differently colored bars, and you can navigate and zoom using the toolbar at the top of the screen. The screenshot in the figure below shows an example of insertion inside green LCB of the assembled genome.

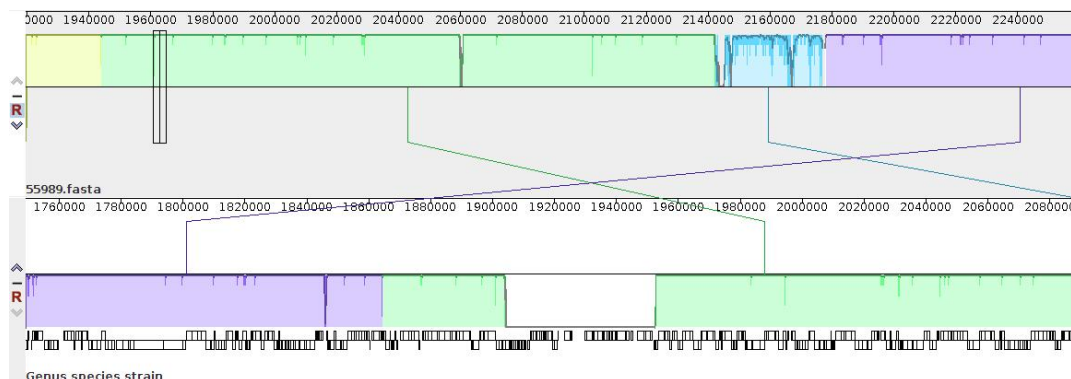


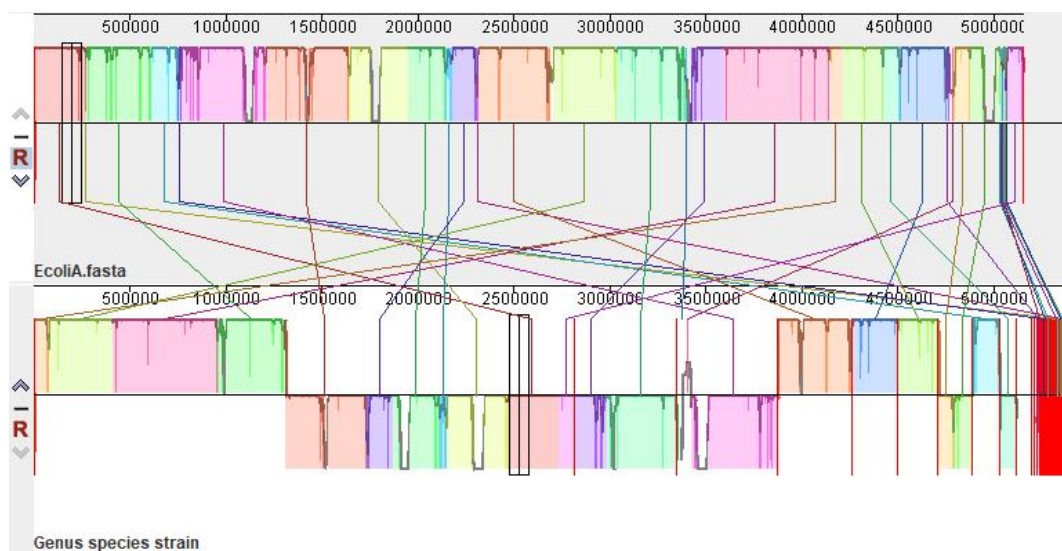
Figure: Screenshot of Mauve visualization of two closely related genomes. The white gap in the lower genome may indicate an insertion into the bacterial chromosome.

Since bacterial genomes undergo rearrangements, various regions can be rearranged, even between two closely related bacterial species. Collinear blocks allow us to visualize these rearrangements with ease.

To compare *E. coli* X with *E. coli* A, first install Mauve (<http://darlinglab.org/mauve/download.html>) on your computer.

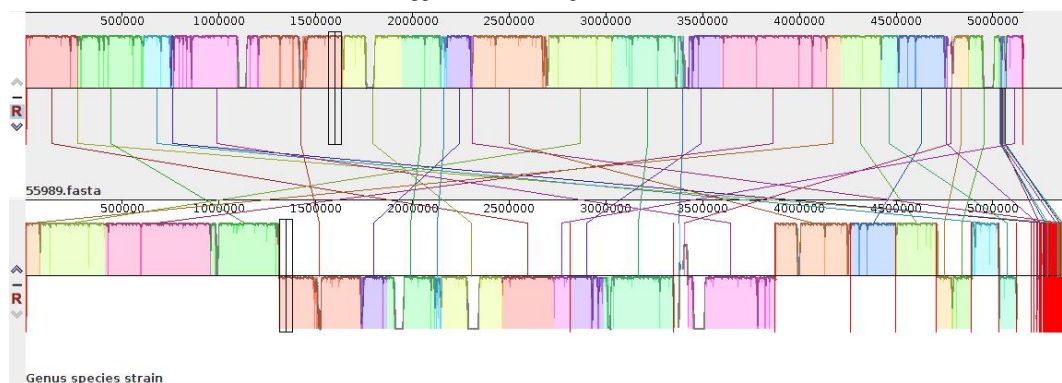
Open "Mauve" and select "File" → "Align with progressiveMauve...". Press "Add sequences" and select *E. coli* A as the reference genome ("EcoliA.fasta"), then the annotated *E. coli* X genome ("scaffolds.gbk"), and start the alignment.

Question 9: Attach a screenshot of the resulting alignment (select "Tools" → "Export" → "Export Image...")



RUBRIC

The learner's submission should resemble the following:

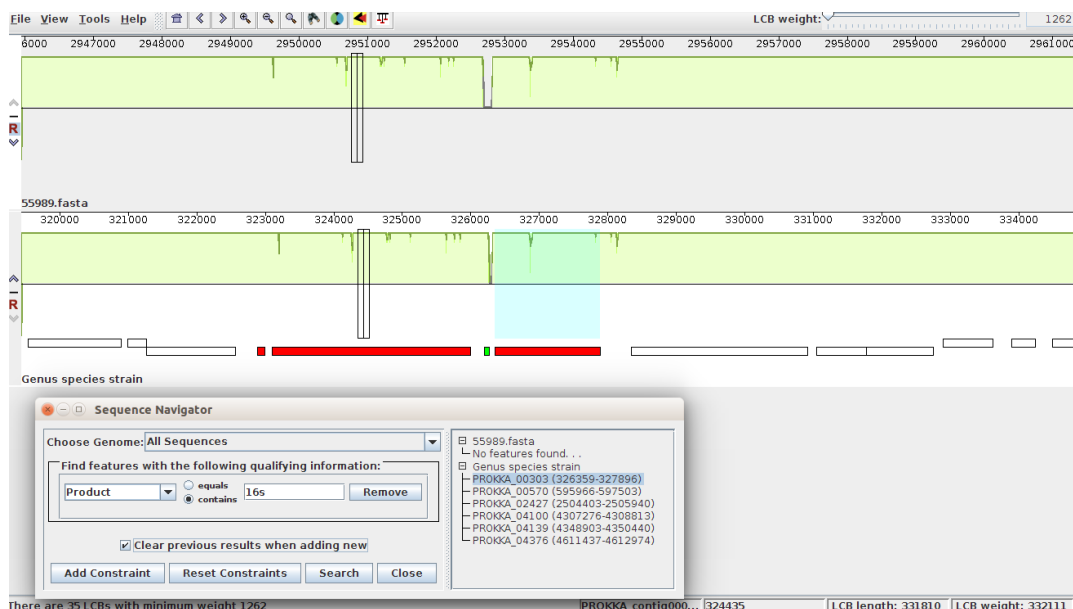


- ☐ 0 pts
Answer is incorrect
- ☐ 1 pt
Answer is correct

PROMPT

We will now analyze the genome-wide alignment of *E. coli* X and *E. coli* A. We are interested in unaligned regions in the scaffolds as putative insertions between *E. coli* A and *E. coli* X to figure out how *E. coli* X acquired the additional virulence factor of causing bleeding. In particular, we will check whether *E. coli* X encodes the Shiga toxins produced by other *E. coli* strains that cause internal bleeding.

You can look for specific genes using Sequence Navigator in Mauve (the icon of the binoculars in the upper line). Select "Product" in the left window and enter the name of the desired gene (or its function) in the right window.



Here is an example of 16s rRNA search in the assembled genome.

Question 10: What are the names of the shiga toxin-related genes that you find?

stxB and stxA

RUBRIC

stxA and stxB

- ☐ 0 pts
Learner did not provide correct gene names
- ☐ 1 pt
Learner provide only one correct gene name
- ☐ 2 pts
Learner provide correct names for both genes

PROMPT

Tracing the source of shiga toxin genes in *E. coli* X

We have discovered that the *E. coli* X genome encodes Shiga-like toxin genes (see Figure below). Now let's figure out how this strain has acquired these weapons.



Figure: Diagram of Shiga toxin in *S. dysenteriae*. The toxin has two subunits encoded by separate genes that are joined by a bond so that they can work together to cleave the host's ribosomes.

We have found two *Stx* genes (269 and 959 bp in length) in the *E. coli* X genome. But how did *E. coli* X acquire these genes? Zoom out in the Mauve window and you will see that the region containing toxin genes is located in a large segment inserted at position 1,176,265 in the *E. coli* A genome.

Indeed, the purple bars flanking this insertion correspond to an aligned locally collinear block shared between the reference *E. coli* A above and *E. coli* X (the insertion itself is represented as whitespace).

The bars below the insertion, representing genes, reveal that this insertion contains many more genes than just the two Shiga toxin genes. Hopefully, further analysis of these genes will help us identify the origin of this insertion.

Move your mouse cursor over genes in this insertion. You will see that most known proteins in the inserted segment are related to phages – *E. coli* X must have therefore gained an additional virulence factor from a phage! In particular, the EHEC *E. coli* A strain obtained a shiga-carrying *Stx* phage that transformed it into the dangerous *E. coli* X strain. The EHEC traits enabled the new strain to

aggressively colonize the mucosa and thereby facilitate absorption of Shiga toxin, which promoted progression to HUS in patients (see FAQ: Classification of pathogenic *E. coli* strains).

Antibiotic resistance and its origin

During the German outbreak of *E. coli* X, some patients needed therapeutic or prophylactic administration of antibiotics. The situation was complicated by the fact that some antibiotics, such as ciprofloxacin, can actually activate stx2-harboring phages, increasing Stx2 production (Bielaszewska et al. 2012), and so additional studies were necessary for these patients in order to choose proper treatment.

Now that we understand the genetic identity of *E. coli* X, we can use the information we have learned to approach this problem.

To search for genes responsible for antibiotic resistance, we will use ResFinder (<https://cge.cbs.dtu.dk/services/ResFinder/>) which specifically searches a database of genes implicated in antibiotics resistance, identifying similarities between the sequenced genome and this database using local alignment.

Visit the ResFinder homepage, then upload the scaffolds.fasta file from the SPAdes output, and in the field "Select Antimicrobial configuration" select "All". For comparison, do the same for the *E. coli* A strain.

Question 11: Which antibiotics are *E. coli* X and *E. coli* A resistant to?

E. coli A: Tetracycline

E. coli X: Aminoglycoside, Beta-lactam,
Sulphonamide, Tetracycline, Trimethoprim

RUBRIC

E. coli X is resistant to aminoglycosides, β -lactams, sulphonamide, tetracycline and trimethoprim. *E. coli* A is resistant to tetracycline only.

- ☐ 0 pts
Learner provide incorrect list for both strains
- ☐ 1 pt
Learner provide incorrect list for one of the strains
- ☐ 2 pts
Both lists are correct

PROMPT

We have discovered that *E. coli* X acquired not only the Shiga toxin but additional antibiotics resistance via HGT! But how?

β -lactam antibiotics, which include penicillin, refer to all antibiotic agents that contain a molecular structure called the **β -lactam ring** (Figure below). These antibiotics stop bacterial growth by inhibiting biosynthesis of the cell wall. However, some bacteria found a defense against these antibiotics by acquiring the β -lactamase enzyme. This enzyme renders a β -lactam antibiotic ineffective by cleaving its β -lactam ring.

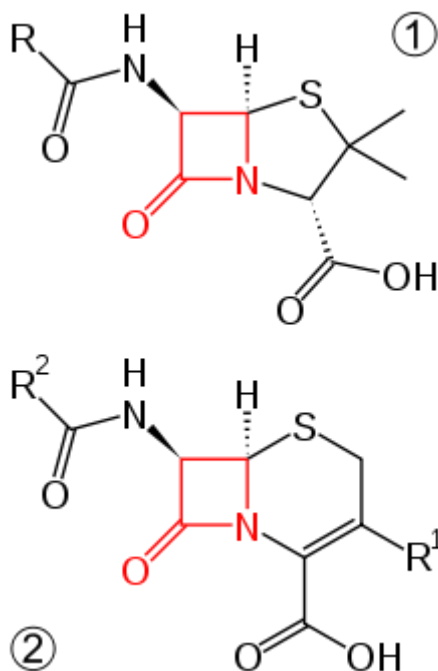


Figure: Structures of penicillin (top) and cephalosporin (bottom). The β -lactam ring is highlighted in red. Courtesy: Fvasconcellos, Wikimedia Commons user.

Let's return to our alignment and search for β -lactamase enzymes, which are usually encoded by genes called **bla**. You can search for these enzymes in the same way that we looked for the toxin genes, by using the Sequence Navigator in Mauve.

Question 12: How many bla genes do you find? How do you think that *E. coli* X obtained these genes?

there are 2 bla genes. Like other bacteria, probably because of HGT (horizontal gene transfer)